HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Matematiikan ja tilastotieteen laitos

Tekijä — Författare — Author
Antti Kuukka

Työn nimi — Arbetets titel — Title

An Application of Ergodic Theory: The Shannon-McMillan-Breiman Theorem

Oppiaine — Läroämne — Subject		
Sovellettu matematiikka		
Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Helmikuu 2013	63 s.

Tiivistelmä — Referat — Abstract

Työn päätavoitteena on todistaa eräs versio informaatioteorian kuuluisasta tuloksesta, Shannon-McMillan-Breiman -teoreemasta. Koska tämä lause liittyy ergodisiin stokastisiin prosesseihin ja sen todistuksessa tarvitaan ergoditeorian tuloksia, olen ottanut ergoditeorian toiseksi tasavertaiseksi pääaiheeksi Shannon-McMillan-Breiman -teoreeman rinnalle.

Lukijalta ei oleteta esitietovaatimuksena ergoditeoriasta tai informaatioteorista mitään. Vaikka mittateoreettisen todennäköisyysteorian ja Markovin ketjujen perusteet oletetaan tunnetuiksi, kappaleessa 0 käydään läpi tiettyjä todennäköisyysteorian osa-alueita, jotka eivät alan peruskursseille mahdu. Näitä ovat esimerkiksi ääretönulotteisiin jonoavaruuksiin konstruoitavat sigma-algebrat, diskreettiaikainen martingaalikonvergenssiteoria, Radon-Nikodym - lauseeseen perustuva ehdollinen todennäköisyys, sekä tasainen integroituvuus.

Kappale 1 käsittelee ergoditeoriaa. Koska ergoditeoria on hyvin laaja matematiikan osa-alue, tämä kappale on aiheeseen vain lyhyt johdatus. Eräs alan tärkeimmistä tuloksista, Birkhoffin ergodilause kuitenkin esitetään kappaleessa todistuksineen. Ergoditeorian voidaan ajatella olevan matematiikan osa-alue, joka tutkii ilmiöiden keskimääräistä käyttäytymistä. Birkhoffin ergodilauseesta esimerkiksi seuraa lähes välittömästi vahva suurten lukujen laki sekä se, että pelkistymättömän ja jaksottoman Markovin ketjun tietyssä tilassa viettämä suhteellinen aika on asymptoottisesti sama kuin tasapainojakauman kyseistä tilaa koskeva pistetodennäköisyys. Kappaleen lopuksi määritellään käsite ergodinen stokastinen prosessi, joita Shannon-McMillan-Breiman -lauseen väite koskee.

Kappaleen 2 aiheena on informaatioteoria ja Shannon-McMillan-Breiman -lause. Koska lukijalta ei oleteta minkäänlaista etukäteistietoa informaatioteoriasta, kappale alkaa johdatuksella informaatioteoriaan. Johdatuksessa keskitytään ennen kaikkea diskreettien satunnaismuuttujien entropiaan, jota koskevia lauseita ja aputuloksia esitetään runsaasti. Tämän jälkeen satunnaismuuttujien entropiasta siirrytään stokastisten prosessien entropian tutkimiseen, jonka jälkeen on mahdollista esittää käsite *asymptoottinen tasapartitiointiominaisuus*. Karkeasti ottaen voidaan sanoa, että diskreetillä stokastisella prosessilla on asymptoottinen tasapartitiointiominaisuus, mikäli melkein kaikki sen realisaatiot kuuluvat alkioiden lukumäärältään pieneen, mutta todennäköisyysmassaltaan suureen *tyypillisten jonojen joukkoon*, joka käsitteenä määritellään kappaleessa tarkasti. Esimerkiksi tasapainoista kolikkoa heitettäessä tyypillisiä jonoja ovat sellaiset realisaatiot, joissa noin puolet heitoista on klaavoja. Paljastuu, että realisaatioiden jakamisella tyypillisiin ja ei-tyypillisiin jonoihin on mielenkiintoisia sovelluksia, kuten tiedon tiivistäminen.

Shannon-McMillan-Breiman -lause sanoo, että stationaarisella, ergodisella stokastisella prosessilla, jolla on äärellinen tila-avaruus, on asymptoottinen tasapartitiointiominaisuus. Kappaleen 2 pituudesta huomattava osuus menee tämän tuloksen todistamiseen. Todistus on monivaiheinen, ja se hyödyntää runsaasti erilaisia tuloksia informaatioteorian ulkopuolelta sekä punoo yhteen työssä aiemmin johdetut tulokset. Birkhoffin ergodilause ja kappaleessa 0 esitettävä Levyn martingaalikonvergenssilause ovat erityisen tärkeässä roolissa.

Avainsanat — Nyckelord — Keywords
Informaatioteoria, ergoditeoria

Säilytyspaikka — Förvaringsställe — Where deposited
Kumpulan tiedekirjasto

Muita tietoja — Övriga uppgifter — Additional information

University of Helsinki
Department of Mathematics & Statistics
Applied mathematics

Master's Thesis

An Application of Ergodic Theory: The Shannon-McMillan-Breiman Theorem

Author:
Antti Kuukka

Supervisor:
Esa Nummelin

February 11, 2013

Contents

Preface

The subject of this Master's Thesis is Shannon-McMillan-Breiman theorem, a famous and important result in information theory. Since the theorem is a statement about *ergodic* stochastic processes and its proof utilises Birkhoff's ergodic theorem, a whole chapter has been devoted to ergodic theory.

Ergodic theory has developed into a large branch of mathematics, and so the Chapter 1 is only a brief glance at the subject. Nevertheless, we will prove one of the most important theorems in ergodic theory, the before-mentioned Birkhoff's ergodic theorem. This theorem is a strong statement about the average behaviour of certain stochastic processes (or *dynamical systems*), and it can be seen as a generalisation of the Strong Law of Large Numbers.

Chapter 2 discusses information theory and the Shannon-McMillan-Breiman theorem. No previous knowledge about information theory is assumed, and therefore the chapter starts with an introduction to information theory. All fundamental definitions and theorems concerning the entropy of discrete random variables are provided. After this introduction, we study the entropy of stochastic processes, which in turn leads us to the Asymptotic Equipartition Property (the AEP). Informally, a stochastic process has the AEP if almost all sample paths belong to a rather thin set, called the set of *typical sequences*, which despite having few elements contains most of the probability mass. Then we prove that independent and identically distributed processes have the AEP, and consider its consequences and applications such as data compression. After this, we present the Shannon-McMillan-Theorem which states that stationary, ergodic processes with finite state space have the AEP. The rest of the thesis is then devoted to the rather long, but interesting proof of the theorem.

The reader is assumed to have basic knowledge about measure-theoretic probability theory. Familiarity with Markov chains, which form an important class of stationary, ergodic processes, is also assumed. They will appear in numerous examples throughout the text. However, my aim has been to make this text as self-contained as possible, and therefore a preliminary Chapter 0 is included. Topics discussed in this chapter include infinite dimensional product spaces and sigma-algebras, discrete-time stochastic processes, conditional probability and discrete-time martingale convergence theory.

Nearly all theorems and lemmas presented in this Master's Thesis are also proved. Most notable exceptions are the Kolmogorov Extension Theorem, the π-λ theorem and the Radon-Nikodym theorem. The proofs had to be omitted in order to keep Chapter 0 reasonably short.

Frequently Used Notation

I. Sets and Spaces

(1) $\emptyset$ is the empty set.
(2) $\mathbb{N}$ is the set of natural numbers $1, 2, 3, \ldots$.
(3) $\mathbb{N}_0$ is the set of consisting of zero and the natural numbers $1, 2, 3, \ldots$.
(4) $\mathbb{Z}$ is the set of integers.
(5) $\mathbb{R}$ is the set of real numbers.
(6) $\mathbb{R}^n$ is the n-dimensional real space.
(7) $\overline{\mathbb{R}}$ is the extended set of real numbers, that is, $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty, -\infty\}$.
(8) $\mathscr{B}(X)$ is the collection of Borel subsets of topological space X.
(9) If X is any set, then $\mathcal{P}(X)$ is the power set of X, that is, the collection of all subsets of X.
(10) The number of elements in X is denoted by $|X|$. If X is infinite, then $|X| = \infty$.

II. σ-algebras

(1) $\sigma(\mathscr{C})$ is the σ-algebra generated by $\mathscr{C}$, that is, the intersection of all σ-algebras that contain $\mathscr{C}$.
(2) $\prod_{k=1}^{n} \mathscr{F}_k$, or $\mathscr{F}_1 \times \mathscr{F}_2 \times \cdots \times \mathscr{F}_n$, is the product σ-algebra of σ-algebras $\mathscr{F}_1, \mathscr{F}_2, \ldots, \mathscr{F}_n$. It may be written as $\mathscr{F}^n$ if $\mathscr{F}_k = \mathscr{F}$ for $k \leq n$.

III. Limits

(1) If $A, A_1, A_2, \ldots$ are subsets of some set Ω, then $A_n \uparrow A$ means that $A_1 \subset A_2 \subset A_3 \subset \ldots$ and $\bigcup_{i=1}^{\infty} A_i = A$. Similarly, $A_n \downarrow A$ means that $A_1 \supset A_2 \supset A_3 \supset \ldots$ and $\bigcap_{i=1}^{\infty} A_i = A$.
(2) If $x, x_1, x_2, \ldots$ belong to $\overline{\mathbb{R}}$ (or $\mathbb{R}$), then $x_n \uparrow x$ means that the x_n form an increasing sequence and $\lim_{n\to\infty} x_n = x$.
(3) If $f, f_1, f_2, \ldots$ are functions from Ω to $\overline{\mathbb{R}}$ (or $\mathbb{R}$), then $f_n \uparrow f$ means that $f_n(\omega) \uparrow f(\omega)$ holds for each $\omega \in \Omega$.

IV. Functions

(1) I_A is the indicator function of set A, that is, $I_A(\omega) = 1$ for $\omega \in A$ and 0 for $\omega \in A^c$.
(2) If $f : \Omega \to \overline{\mathbb{R}}$ is a function, then $f^+ = \max\{f, 0\}$ is the positive part and $f^- = -\min\{f, 0\}$ is the negative part of f.

Other Comments on Notation

If $(\Omega_1, \mathscr{F}_1)$ and $(\Omega_2, \mathscr{F}_2)$ are measurable spaces, we say that a function $f : \Omega_1 \to \Omega_2$ is measurable $\mathscr{F}_1/\mathscr{F}_2$ if we have $f^{-1}A \in \mathscr{F}_1$ for all $A \in \mathscr{F}_2$. If $\Omega_2 = \mathbb{R}^n$ and $\mathscr{F}_2 = \mathscr{B}(\mathbb{R}^n)$, we may indicate that f is measurable $\mathscr{F}_1/\mathscr{F}_2$ by saying that f is *Borel measurable.*

If $X : \Omega_1 \to \Omega_2$ is measurable $\mathscr{F}_1/\mathscr{F}_2$ and $\mathscr{F}_1$ is equipped with a probability measure P, then X is called a *random variable.* Thus, since $(\Omega_2, \mathscr{F}_2)$ is not necessarily $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, we do not assume that random variables are real-valued functions, and therefore there is no strict distinction between random variables and vectors. But the term random vector may be used when convenient, especially when $X_1, X_2, \ldots, X_n$ are random variables and we want to treat them as a single object $(X_1, X_2, \ldots, X_n)$.

The *distribution* of random variable X is the probability measure P_X on $\mathscr{F}_2$ defined by

$$P_X(A) = P(X \in A),\ A \in \mathscr{F}_2.$$

If $(\Omega, \mathscr{F}, P)$ is a probability space, then the space of all real-valued integrable random variables (that is, functions f for which $E[|f|] < \infty$) is denoted by $L^1(\Omega)$. If $Y, Y_1, Y_2, \ldots \in L^1(\Omega)$, then we may say that the Y_n converge to Y in L^1 if $E[|Y_n - Y|] \to 0$ as $n \to \infty$. And if $\sup_{n\in\mathbb{N}} E[|Y_n|] < \infty$, then we say that the Y_n are *bounded in* L^1.

Chapter 0

Preliminaries

In this chapter we discuss certain topics in probability theory that are essential prerequisities for the later chapters.

0.1 The π-λ theorem and uniqueness of probability measures

Suppose that $\mathscr{F}$ is a σ-algebra, and we want to prove that some property holds for all $A \in \mathscr{F}$. For instance, we may want to prove that probability measures P and Q agree on $\mathscr{F}$, that is, $P(A) = Q(A)$ for all $A \in \mathscr{F}$. Although it may be difficult to check directly that the property truly holds for all $A \in \mathscr{F}$, it often suffices to check that the property holds in a collection of subsets that generates $\mathscr{F}$. As we will see, this is possible if the generating set is a π-system, and the class of sets for which the property holds is a λ-system.

Definition 0.1. Suppose that Ω is a nonempty set and $\mathscr{P} \subset \mathcal{P}(\Omega)$. Then $\mathscr{P}$ is called a *π-system* if $A, B \in \mathscr{P}$ implies that $A \cap B \in \mathscr{P}$, that is, $\mathscr{P}$ is closed under finite intersections.

Definition 0.2. Suppose that Ω is a nonempty set and $\mathscr{L} \subset \mathcal{P}(\Omega)$. Then $\mathscr{L}$ is a λ-system if the following conditions are met:

(1) $\Omega \in \mathscr{L}$;
(2) $A \in \mathscr{L}$ implies $A^c \in \mathscr{L}$;
(3) if $A_1, A_2, \ldots \in \mathscr{L}$ are disjoint, then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{L}$.

Remark 0.3. It is clear that σ-algebras are λ-systems, but the converse is not true (in a four-point space Ω, let $\mathcal{L}$ consist of $\Omega, \emptyset$ and the six two-point sets).

We may now present the extremely useful π-λ theorem, which will be applied numerous times in this Master's Thesis. Its rather technical proof is omitted.

Theorem 0.4. *(The π-λ theorem) If $\mathscr{P}$ is a π-system and $\mathscr{L}$ is a λ-system, then $\mathscr{P} \subset \mathscr{L}$ implies that $\sigma(\mathcal{P}) \subset \mathcal{L}$.*

Proof. See [3, p. 42] . □

To illustrate how the π-λ theorem is used in practice, we will now prove an importantant uniqueness theorem for probability measures.

Theorem 0.5. *Suppose that $\mathscr{P}$ is a π-system, and P and Q are probability measures on $\sigma(\mathscr{P})$. If P and Q agree on $\mathscr{P}$, then $P = Q$.*

Proof. Let $\mathscr{L} = \{A \in \sigma(\mathscr{P}) : P(A) = Q(A)\}$. If we can show that $\mathscr{L}$ is a λ-system, then by hypothesis $\mathscr{P} \subset \mathscr{L}$ and the π-λ theorem implies that $\sigma(\mathscr{P}) \subset \mathscr{L}$, that is, we have $P(A) = Q(A)$ for all $A \in \sigma(\mathscr{P})$.

Of course $\Omega \in \mathscr{L}$, and if $A \in \mathscr{L}$, then $P(A^c) = 1 - P(A) = 1 - Q(A) = Q(A^c)$, and thus $A^c \in \mathscr{L}$. If $A_1, A_2, \ldots$ are disjoint $\mathscr{L}$-sets, then $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} Q(A_n) = Q(\bigcup_{n=1}^{\infty} A_n)$, which implies that $\bigcup_{n=1}^{\infty} A_n \in \mathscr{L}$. Therefore, $\mathscr{L}$ is a λ-system. □

0.2 Infinite product spaces and σ-algebras

Recall from probability theory that if $(\Omega_1, \mathscr{F}_1), (\Omega_2, \mathscr{F}_2), \ldots, (\Omega_n, \mathscr{F}_n)$ are measurable spaces, then the product space $\Omega_1 \times \Omega_2 \times \cdots \Omega_n$ may be equipped with the σ-algebra $\prod_{k=1}^{n} \mathscr{F}_k$ generated by measurable rectangles, that is, sets of form $F_1 \times F_2 \times \cdots \times F_n$ with $F_i \in \mathscr{F}_i$ for $1 \leq i \leq n$. Morever, if each $(\Omega_k, \mathscr{F}_k)$ is equipped with a probability measure P_k, then there exists a unique probability measure P on $\prod_{k=1}^{n} \mathscr{F}_k$, called the product measure of $P_1, P_2, \ldots, P_n$, such that

$$P(F_1 \times F_2 \times \cdots \times F_n) = P_1(F_1)P_2(F_2) \cdots P_n(F_n)$$

for all measurable rectangles. Our aim is to extend this idea to infinite products of probability spaces.

Suppose that T is any ordered set and $(\Omega_k, \mathscr{F}_k)_{k \in T}$ is a collection of measurable spaces. Let $\prod_{k \in T} \Omega_k$ be the product space formed by the sets Ω_k. For example, if $T = \mathbb{Z}$, then the elements of $\prod_{k \in T} \Omega_k$ are sequences $\omega = (\ldots, \omega_{-1}, \omega_0, \omega_1, \ldots)$ such that $\omega_k \in \Omega_k$ for all $k \in \mathbb{Z}$. If $T = \mathbb{N}$, then the product space consists of sequences $\omega = (\omega_1, \omega_2, \ldots)$ with $\omega_k \in \Omega_k$ for all $k \in \mathbb{N}$. If $\Omega_k = \Omega$ for all $k \in T$, then we may write $\prod_{k \in T} \Omega_k = \Omega^T$; if $T = \mathbb{N}$, it is customary to write Ω^T as Ω^∞. We want to equip $\prod_{k \in T} \Omega_k$ with a σ-algebra.

Let $k_1 < k_2 < \ldots < k_n \in T$ and $B \subset \prod_{i=1}^n \Omega_{k_i}$. Define

$$\mathcal{C}(B) = \{\omega : (\omega_{k_1}, \omega_{k_2}, \ldots, \omega_{k_n}) \in B\}.$$

If $B \in \prod_{i=1}^n \mathscr{F}_{k_i}$, then $\mathcal{C}(B)$ is called a *cylinder* with *base* B at $(k_1, k_2, \ldots, k_n)$. The cylinder is called a *measurable rectangle* if B is of form $B_1 \times \cdots \times B_n$ with $B_i \in \mathscr{F}_{k_i}$ for all $1 \leq i \leq n$.

Let $\prod_{k\in T} \mathscr{F}_k$ be the σ-algebra generated by the cylinder sets. If $\mathscr{F}_k = \mathscr{F}$ for all $k \in T$, then we may write $\prod_{k\in T} \mathscr{F}_k = \mathscr{F}^T$ (if $T = \mathbb{N}$, then $\mathscr{F}^T$ may also be written as $\mathscr{F}^\infty$). By the following lemma, $\prod_{k\in T} \mathscr{F}_k$ is also generated by measurable rectangles.

Lemma 0.6. *The σ-algebras generated by cylinder sets and measurable rectangles coincide with each other.*

Proof. Since measurable rectangles are cylinders, it is clear that it is enough to show that the σ-algebra $\mathscr{F}_{MR}$ generated by the measurable rectangles contains the cylinder sets. For each $n \in \mathbb{N}$ and $k_1 < k_2 < \ldots < k_n \in T$, put

$$\mathscr{C}_{k_1,k_2,\ldots,k_n} = \{A \subset \prod_{i=1}^n \Omega_{k_i} : \mathcal{C}(A) \in \mathscr{F}_{MR}\}.$$

It is easy to check that $\mathscr{C}_{k_1,k_2,\ldots,k_n}$ is a σ-algebra. But sets of form $F_1 \times F_2 \times \cdots \times F_n, F_i \in \mathscr{F}_i$, clearly belong to $\mathscr{C}_{k_1,k_2,\ldots,k_n}$. These sets generate the σ-algebra $\prod_{i=1}^n \mathscr{F}_{k_i}$, which implies that all cylinder sets with bases at $(k_1, k_2, \ldots, k_n)$ belong to $\mathscr{F}_{MR}$. □

Observe that cylinders do not have unique bases. For example, if $B \in \mathscr{F}_1$, then $\mathcal{C}(B) = \mathcal{C}(B \times \Omega_2)$. This idea is formalised in the next lemma.

Lemma 0.7. *Let $\mathcal{C}(B)$ be a cylinder with base B at $(k_1, k_2, \ldots, k_n)$. Suppose that $(k_1, k_2, \ldots, k_n) \subset (j_1, j_2, \ldots, j_m)$, $j_1 < j_2 < \ldots < j_m$. Then there exists a base set B' at $(j_1, j_2, \ldots, j_m)$ such that $\mathcal{C}(B) = \mathcal{C}(B')$.*

Proof. If $B' = \{\omega \in \prod_{k=1}^m \Omega_{j_k} : (\omega_{k_1}, \omega_{k_2}, \ldots, \omega_{k_n}) \in B\}$, then clearly $\mathcal{C}(B) = \mathcal{C}(B')$. Therefore, it suffices to show that $B' \in \prod_{k=1}^m \mathscr{F}_{j_k}$.

Let $\mathscr{C}$ be the class of sets C in $\prod_{i=1}^n \mathscr{F}_{k_i}$ for which $\{\omega \in \prod_{k=1}^m \Omega_{j_k} : (\omega_{k_1}, \omega_{k_2}, \ldots, \omega_{k_n}) \in C\}$ belongs to $\prod_{k=1}^m \mathscr{F}_{j_k}$. It is easy to check that $\mathscr{C}$ is a σ-algebra. Since measurable rectangles (which generate $\prod_{i=1}^n \mathscr{F}_{k_i}$) belong to $\mathscr{C}$, it follows that $\prod_{i=1}^n \mathscr{F}_{k_i} \subset \mathscr{C}$ and thus $B' \in \mathscr{C}$. □

Thus if $\mathcal{C}(B_1)$ and $\mathcal{C}(B_2)$ are two cylinders with bases B_1 and B_2, we may assume that the coordinates of the base sets are the same. Now the following result is easy to prove:

Lemma 0.8. *The cylinder sets form a π-system, and so do the measurable rectangles.*

Proof. Let $\mathcal{C}(B_1)$ and $\mathcal{C}(B_2)$ be two cylinders with bases B_1 and B_2 at $(k_1, k_2, \ldots, k_n)$. Then

$$\begin{aligned}\mathcal{C}(B_1) \cap \mathcal{C}(B_2) &= \{\omega : (\omega_{k_1}, \ldots, \omega_{k_n}) \in B_1\} \cap \{\omega : (\omega_{k_1}, \ldots, \omega_{k_n}) \in B_2\} \\ &= \{\omega : (\omega_{k_1}, \ldots, \omega_{k_n}) \in B_1 \cap B_2\} = \mathcal{C}(B_1 \cap B_2).\end{aligned}$$

Since $B_1 \cap B_2 \in \prod_{i=1}^n \mathscr{F}_{k_i}$, $\mathcal{C}(B_1) \cap \mathcal{C}(B_2)$ is a cylinder with base $B_1 \cap B_2$ at $(k_1, k_2, \ldots, k_n)$. This proves that the cylinder sets form a π-system.

If $\mathcal{C}(B_1)$ and $\mathcal{C}(B_2)$ are measurable rectangles, then the intersection $B_1 \cap B_2$ is again a cartesian product, which implies that $\mathcal{C}(B_1 \cap B_2)$ is a measurable rectangle. Therefore, the measurable rectangles form a π-system as well. □

In the next section, we will construct probability measures on product σ-algebras.

0.3 Stochastic processes

A *stochastic process* is a collection $(X_t)_{t \in T}$ of random variables defined on some probability space $(\Omega, \mathscr{F}, P)$. The random variables take values in a second measurable space $(S, \mathscr{S})$ called the *state space*. The *parameter set* T is usually $[0, \infty)$ (a continuous time process), $\mathbb{Z}$ or $\mathbb{N}$ (discrete time processes). We note that

- for each $t \in T$, the function $\omega \mapsto X_t(\omega)$ is measurable $\mathscr{F}/\mathscr{S}$,
- for a fixed sample point $\omega \in \Omega$, the function $t \mapsto X_t(\omega)$ is called the *sample path* of the process associated with ω.

From here on, we will only discuss discrete time processes with $T = \mathbb{N}$ or $T = \mathbb{Z}$. The state space will usually be $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ or some countable set S in which case we may take $\mathcal{P}(S)$ as the σ-algebra.

Definition 0.9. If $k_1 < k_2 < \ldots < k_n$ and $k_i \in T$ for each $i \leq n$, then the *marginal distribution* of $X_{k_1}, X_{k_2}, \ldots, X_{k_n}$ is the probability distribution

$$P_{X_{k_1}, X_{k_2}, \ldots, X_{k_n}}(A) = P[(X_{k_1}, X_{k_2}, \ldots, X_{k_n}) \in A], \ A \in \mathscr{S}^n.$$

Observe that if $T = \mathbb{N}$ and $A' = \{s \in S^{k_n} : (s_{k_1}, s_{k_2}, \ldots, s_{k_n}) \in A\}$, then

$$P_{X_{k_1}, X_{k_2}, \ldots, X_{k_n}}(A) = P[(X_1, X_2, \ldots, X_{k_n}) \in A'] = P_{X_1, X_2 \ldots, X_{k_n}}(A'),$$

and we conclude that all the marginal distributions are determined by the marginal distributions of $X_1, X_2, \ldots, X_n$, $n \in \mathbb{N}$. Similarly, if $T = \mathbb{Z}$, the marginal distributions are determined by the marginal distributions of $X_{-n}, \ldots, X_0, \ldots, X_n$, $n \in \mathbb{N}$.

The concept of stationarity will play an important role later in this text. Intuitively, it means that the probability structure of the process is independent of time.

Definition 0.10. Let $(X_t)_{t\in T}$ be a stochastic process with $T = \mathbb{N}$ or $T = \mathbb{Z}$. We say that the process is *stationary* if for any $n, m \in \mathbb{N}$ and $k_1 < k_2 < \ldots < k_n \in T$, the marginal distribution of $X_{k_1}, X_{k_2}, \ldots, X_{k_n}$ equals the marginal distribution of $X_{k_1+m}, X_{k_2+m}, \ldots, X_{k_n+m}$.

For example, a sequence of independent, identically distributed random variables is clearly a stationary process. Another good example of a stationary process is an aperiodic, irreducible Markov chain: if the initial distribution of the Markov chain equals its stationary distribution, then the process is stationary.

Distribution of a stochastic process

Suppose that $X = (X_k)_{k\in T}$ is a stochastic process defined on $(\Omega, \mathscr{F}, P)$. The process defines a function $X : \Omega \to S^T$ by

$$(X(\omega))_t = X_t(\omega), \quad t \in T.$$

For example, if $T = \mathbb{N}$, then $X(\omega) = (X_1(\omega), X_2(\omega), \ldots)$. Let us equip the space S^T with the σ-algebra $\mathscr{S}^T$ generated by the cylinder sets. We want to show that X is measurable $\mathscr{F}/\mathscr{S}^T$.

Lemma 0.11. *Suppose that $(\Omega, \mathscr{F})$ is a measurable space and f is a function from Ω to Ω'. Let $\mathscr{G}$ be a collection of subsets of Ω'. If $f^{-1}G \in \mathscr{F}$ for all $G \in \mathscr{G}$, then f is measurable $\mathscr{F}/\sigma(\mathscr{G})$.*

Proof. Let $\mathscr{C} = \{C \subset \Omega' : f^{-1}C \in \mathscr{F}\}$. It is easy to see that $\mathscr{C}$ is a σ-algebra. But $\mathscr{G} \subset \mathscr{C}$ and therefore $\sigma(\mathscr{G}) \subset \mathscr{C}$. This proves that f is measurable $\mathscr{F}/\sigma(\mathscr{G})$. □

Now let A be a one-dimensional cylinder set, that is, $A = \{s \in S^T : s_k \in A'\}$ for some $k \in T$ and $A' \in \mathscr{S}$. Since measurable rectangles can be written as intersections of one-dimensional cylinders, the σ-algebra generated by the one-dimensional cylinder sets coincides with $\mathscr{S}^T$. And since

$$X^{-1}A = \{\omega : X(\omega) \in A\} = \{\omega : X_k(\omega) \in A'\} = X_k^{-1}A' \in \mathscr{F},$$

the previous lemma implies that X is measurable $\mathscr{F}/\mathscr{S}^T$. Hence, stochastic processes are random variables taking values in S^T. We may now define the distribution of a stochastic process:

Definition 0.12. The *distribution* of a stochastic process X is the probability distribution P_X on $\mathscr{S}^T$ defined by the formula $P_X(A) = P(X \in A), A \in \mathscr{S}^T$.

The distribution of a process is determined by its marginal distributions. This is a direct consequence of Theorem 0.5: if $A = \{s \in S^T : (s_{k_1}, s_{k_2}, \ldots, s_{k_n}) \in A'\}$ is a cylinder set, then $P_X(A)$ is determined by the marginal distribution of $X_{k_1}, X_{k_2}, \ldots, X_{k_n}$. Since $\mathscr{S}^T$ is generated by the cylinders, P_X is uniquely determined.

Recall from probability theory that if X is a real-valued random variable and g is Borel measurable, then

$$\int_\Omega g(X)\,\mathrm{d}P = \int_{\mathbb{R}} g\,\mathrm{d}P_X.$$

An analogous formula holds for our generalized random variables such as stochastic processes.

Lemma 0.13. *Suppose that $X : (\Omega_1, \mathscr{F}_1) \to (\Omega_2, \mathscr{F}_2)$ is measurable $\mathscr{F}_1/\mathscr{F}_2$, and $g : \Omega_2 \to \mathbb{R}$ is a Borel measurable function. If $g(X) \in L^1(\Omega)$, then*

$$\int_{\Omega_1} g(X)\,\mathrm{d}P = \int_{\Omega_2} g\,\mathrm{d}P_X.$$

Proof. The general case is proved just like the special case $\Omega_2 = \mathbb{R}$ (start from indicator functions, then use linearity to prove that the equation holds if g is a simple function, etc.). □

Example 0.14. Let $X = (X_k)_{k\in\mathbb{N}}$ be a stochastic process with state space $(S, \mathscr{S})$. If $f : S^\infty \to \mathbb{R}$ is Borel measurable and $f(X_1, X_2, \ldots)$ is integrable, then

$$\int_\Omega f(X_1, X_2, \ldots)\,\mathrm{d}P = \int_{S^\infty} f\,\mathrm{d}P_X.$$

Existence of stochastic processes

We will now discuss the problem of constructing an underlying probability space $(\Omega, \mathscr{F}, P)$ for a given stochastic process. First, we suppose that $T = \mathbb{N}$. In this case, the following theorem is often very convenient to apply.

Theorem 0.15. *Let $(\Omega_k, \mathscr{F}_k)_{k\in\mathbb{N}}$ be an arbitrary collection of measurable spaces, and let $\Omega = \prod_{k=1}^\infty \Omega_k, \mathscr{F} = \prod_{k=1}^\infty \mathscr{F}_k$. Suppose that we are given a probability measure P_1 on $\mathscr{F}_1$, and for each $n \in \mathbb{N}$ and each $(\omega_1, \omega_2, \ldots, \omega_n) \in \prod_{k=1}^n \Omega_k$ we are given a probability measure $P_{\omega_1,\omega_2,\ldots,\omega_n}$ on $\mathscr{F}_{n+1}$. Assume also that for each fixed $A \in \mathscr{F}_{n+1}$, $P_{\omega_1,\omega_2,\ldots,\omega_n}(A)$, considered as a function of $(\omega_1, \omega_2, \ldots, \omega_n)$, is measurable $\prod_{k=1}^n \mathscr{F}_k/\mathscr{B}(\mathbb{R})$.*

If $B \in \prod_{k=1}^n \mathscr{F}_k$, we define

$$P_n(B) = \int_{\Omega_1} \cdots \int_{\Omega_{n-1}} \int_{\Omega_n} I_B(\omega_1, \ldots, \omega_n) P_{\omega_1,\ldots,\omega_{n-1}}(\mathrm{d}\omega_n) P_{\omega_1,\ldots,\omega_{n-2}}(\mathrm{d}\omega_{n-1}) \cdots P_1(\mathrm{d}\omega_1),$$

which is a well-defined probability measure on $\prod_{k=1}^{n} \mathscr{F}_k$.

Then there exists a unique probability measure P *on* $\mathscr{F}$ *such that for any* $n \in \mathbb{N}$ *and* $B \in \prod_{k=1}^{n} \mathscr{F}_k$, $P[\mathcal{C}(B)] = P_n(B)$.

Proof. See [1, p. 114] □

Now we can use this theorem to construct probability spaces on which stochastic processes are defined. Suppose that our process has state space S equipped with a σ-algebra $\mathscr{S}$. Suppose further that we are given a probability measure P_1 on $\mathscr{S}$ and probability measures $P_{\omega_1,\omega_2,\ldots,\omega_n}$ on $\mathscr{S}$ for each $(\omega_1, \omega_2, \ldots, \omega_n) \in S^n$ in such a way that the measurability condition of the theorem is satisfied. Then the theorem provides us a unique probability measure P on $\mathscr{S}^\infty$. Now, put $X_k(\omega) = \omega_k$. We have for all $B \in \mathscr{S}^n$

$$P[(X_1, X_2, \ldots, X_n) \in B] = P(\{\omega : (\omega_1, \omega_2, \ldots, \omega_n) \in B\}) = P_n(B),$$

where

$$P_n(B) = \int_S \cdots \int_S \int_S I_B(\omega_1, \ldots, \omega_n) P_{\omega_1,\ldots,\omega_{n-1}}(\mathrm{d}\omega_n) P_{\omega_1,\ldots,\omega_{n-2}}(\mathrm{d}\omega_{n-1}) \cdots P_1(\mathrm{d}\omega_1). \tag{0.16}$$

Thus, $(X_k)_{k\in\mathbb{N}}$ is a stochastic process with marginal distributions defined by (0.16).

We shall now use Theorem 0.15 to construct stochastic processes.

1. (Independent sequences) Suppose that for each $k = 1, 2, 3, \ldots$ we are given an arbitrary probability space $(\Omega_k, \mathscr{F}_k, P_k)$. Let $\Omega = \prod_{k=1}^{\infty} \Omega_k$ and $\mathscr{F} = \prod_{k=1}^{\infty} \mathscr{F}_k$. Then there exists a unique probability measure P on $\mathscr{F}$ such that

$$P(\{\omega \in \Omega : \omega_1 \in A_1, \ldots, \omega_n \in A_n\}) = \prod_{k=1}^{n} P_k(A_k) \tag{0.17}$$

for all $n \geq 1$ and all $A_k \in \mathscr{F}_k, k \leq n$. To see this, put $P_{\omega_1,\omega_2,\ldots,\omega_n}(A) = P_{n+1}(A)$, $A \in \mathscr{F}_{n+1}$, and apply the theorem. Then

$$\begin{aligned} P[\mathcal{C}(A_1 \times \cdots \times A_n)] &= \int_{\Omega_1} \cdots \int_{\Omega_n} I_{A_1\times\cdots\times A_n}(\omega_1, \ldots, \omega_n) P_{\omega_1,\omega_2,\ldots,\omega_{n-1}}(\mathrm{d}\omega_n) \cdots P_1(\mathrm{d}\omega_1) \\ &= \int_{\Omega_1} \cdots \int_{\Omega_n} I_{A_1}(\omega_1) \cdots I_{A_n}(\omega_n) P_{\omega_1,\omega_2,\ldots,\omega_{n-1}}(\mathrm{d}\omega_n) \cdots P_1(\mathrm{d}\omega_1) \\ &= \int_{\Omega_1} I_{A_1} \,\mathrm{d}P_1 \int_{\Omega_2} I_{A_2} \,\mathrm{d}P_2 \cdots \int_{\Omega_n} I_{A_n} \,\mathrm{d}P_n = \prod_{k=1}^{n} P_k(A_k). \end{aligned}$$

If Q is any other probability measure on $\mathscr{F}$ with this property, then it agrees with P on the π-system formed by the measurable rectangles. By Theorem 0.5, $P = Q$,

and thus the probability measure given by Theorem 0.15 is the only one with property (0.17). If $X_k(\omega) = \omega_k$, then we obtain a sequence of random variables $X_1, X_2, \ldots$ such that the random variables are independent and $P(X_k \in A) = P_k(A)$.

2. (Infinite fair coin tossing) This is a simple special case of the previous example. Put $\Omega_k = S = \{0, 1\}$, $\mathscr{F}_k = \mathscr{S} = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$ and $P_k(\{0\}) = P_k(\{1\}) = 0.5$ for all k. Then Ω is the space of all infinite sequences consisting of zeros and ones, and if $X_k(\omega) = \omega_k$, then $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{k=1}^{n} P_k(\{x_k\}) = 0.5^n$ for each n.
3. (Markov chains) Let us next consider a Markov chain with finite state space S, initial distribution u and transition probabilities $p(i, j)$, $i, j \in S$. Put $\Omega_k = S$, $\mathscr{F}_k = \mathcal{P}(S)$, $P_1 = u$ and $P_{\omega_1,\omega_2,\ldots,\omega_{n-1}}(\omega_n) = p(\omega_{n-1}, \omega_n)$. Applying the theorem and putting $X_k(\omega) = \omega_k, \omega \in S^\infty$, we obtain a stochastic process with marginal distributions given by

$$\begin{aligned}
P(X_1 = i_1, \ldots, X_n = i_n) &= P[\omega : (\omega_1, \omega_2, \ldots, \omega_n) = (i_1, i_2, \ldots, i_n)] \\
&= \int_S \cdots \int_S I_{\{i_1\}\times\cdots\times\{i_n\}} dP_{\omega_1,\ldots,\omega_{n-1}} \cdots dP_1 \\
&= \int_{\{i_1\}} \cdots \int_{\{i_{n-1}\}} \int_{\{i_n\}} P_{\omega_1,\ldots,\omega_{n-1}}(\mathrm{d}\omega_n) P_{\omega_1,\ldots,\omega_{n-2}}(\mathrm{d}\omega_{n-1}) \cdots P_1(\mathrm{d}\omega_1) \\
&= \int_{\{i_1\}} \cdots \int_{\{i_{n-1}\}} p(\omega_{n-1}, i_n) P_{\omega_1,\ldots,\omega_{n-2}}(\mathrm{d}\omega_{n-1}) \cdots P_1(\mathrm{d}\omega_1) \\
&= \int_{\{i_1\}} \cdots \int_{\{i_{n-2}\}} p(\omega_{n-2}, i_{n-1}) p(i_{n-1}, i_n) P_{\omega_1,\ldots,\omega_{n-3}}(\mathrm{d}\omega_{n-2}) \cdots P_1(\mathrm{d}\omega_1) \\
&= \cdots = u(i_1) p(i_1, i_2) \cdots p(i_{n-1}, i_n).
\end{aligned}$$

Now if $P(X_1 = i_1, \ldots, , X_n = i_n) > 0$, we have

$$\begin{aligned}
P(X_{n+1} = j | X_1 = i_1, \ldots, , X_n = i_n) &= \frac{P(X_1 = i_1, \ldots, , X_n = i_n, X_{n+1} = j)}{P(X_1 = i_1, \ldots, , X_n = i_n)} \\
&= \frac{u(i_1) p(i_1, i_2) \cdots p(i_{n-1}, i_n) p(i_n, j)}{u(i_1) p(i_1, i_2) \cdots p(i_{n-1}, i_n)} = p(i_n, j).
\end{aligned}$$

Therefore, the process $(X_k)_{k\in\mathbb{N}}$ is a Markov chain with finite state space S, initial distribution u and transition probabilities $p(i, j)$, $i, j \in S$.

Kolmogorov extension theorem

Theorem 0.15 assumes that the process has some initial probability distribution P_1. But what if the parameter set T is $\mathbb{Z}$? Then the process has no initial point and Theorem 0.15 can not be applied. In this case, we can use the famous Kolmogorov extension theorem.

Suppose that for all $k_1 < k_2 < \ldots < k_n \in T$ and $n \in \mathbb{N}$, we are given a probability measure $P_{k_1,k_2,\ldots,k_n}$ on the product σ-algebra $\prod_{i=1}^n \mathscr{F}_{k_i}$. Suppose further that these probability measures are *consistent* in the sense that if $(k_1, k_2, \ldots, k_n) \subset (j_1, j_2, \ldots, j_m)$ and $B \in \prod_{i=1}^n \mathscr{F}_{k_i}$, then

$$P_{k_1,k_2,\ldots,k_n}(B) = P_{j_1,j_2,\ldots,j_m}(B') \tag{0.18}$$

for

$$B' = \{(\omega_{j_1}, \omega_{j_2}, \ldots, \omega_{j_m}) \in \prod_{i=1}^m \Omega_{j_i} : (\omega_{k_1}, \omega_{k_2}, \ldots, \omega_{k_n}) \in B\}.$$

We may now apply the Kolmogorov extension theorem to construct probability space for a stochastic process that has its marginal distributions determined by the probability measures $P_{k_1,k_2,\ldots,k_n}$. This theorem is a very powerful one, and it actually works even in continuous time.

Theorem 0.19. *(Kolmogorov extension theorem) Let T be an ordered set. Suppose that for all $k \in T$, Ω_k is a complete, separable metric space and $\mathscr{F}_k = \mathscr{B}(\Omega_k)$. Suppose further that for all $n \in \mathbb{N}$ and $k_1 < k_2 < \ldots < k_n \in T$, we are given a probability measure $P_{k_1,k_2,\ldots,k_n}$ on the product σ-algebra $\prod_{i=1}^n \mathscr{F}_{k_i}$ and these probability measures satisfy the consistency condition (0.18). Then there exists a unique probability measure P on $\prod_{k\in T} \mathscr{F}_k$ that agrees with the probabilities assigned to the cylinder sets, i.e. if $B \in \prod_{i=1}^n \mathscr{F}_{k_i}$, then*

$$P[\mathcal{C}(B)] = P_{k_1,k_2,\ldots,k_n}(B).$$

Proof. See [3, p. 483] □

Suppose that S is a complete, separable metric space. Let $\Omega_k = S$ and $\mathscr{F}_k = \mathscr{S} = \mathscr{B}(S)$ for all $k \in \mathbb{Z}$. If we are given consistent probability distributions $P_{k_1,k_2,\ldots,k_n}$ on $\mathscr{S}^n$ for all $n \geq 1$ and $k_1 < k_2 < \ldots < k_n \in \mathbb{Z}$, then we may apply the Kolmogorov extension theorem, and by putting $X_t(\omega) = \omega_t$ we obtain a stochastic process $(X_k)_{k\in\mathbb{Z}}$ with state space $(S, \mathscr{S})$ and marginal distributions

$$\begin{aligned} P[(X_{k_1}, X_{k_2}, \ldots, X_{k_n}) \in B] &= P(\omega : (\omega_{k_1}, \omega_{k_2}, \ldots, \omega_{k_n}) \in B) = P[\mathcal{C}(B)] \\ &= P_{k_1,k_2,\ldots,k_n}(B),\ B \in \mathscr{S}^n. \end{aligned}$$

Thus, by virtue of the Kolmogorov extension theorem, we can construct a probability space for a stochastic process with marginal distributions determined by any consistent collection of probability distributions. The state space is usually $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ or some finite subset of real numbers together with its power set. In the finite case, observe that any finite subset S of real numbers, equipped with the discrete topology, is a complete, separable metric space, and then we have $\mathscr{B}(S) = \mathcal{P}(S)$ since all subsets of S are open sets.

Example 0.20. Let us construct a probability space for a doubly infinite sequence of Bernoulli trials. Let $0 < p < 1$, and let $\Omega_k = S = \{0, 1\}$, $\mathscr{F}_k = \mathscr{S} = \mathcal{P}(S)$ for all $k \in \mathbb{Z}$. Further, let $P'(\{1\}) = p, P'(\{0\}) = 1-p$, and for all $k_1 < k_2 < \ldots < k_n \in \mathbb{Z}$, let $P_{k_1,k_2,\ldots,k_n}$ be the unique n-dimensional product measure $P' \times P' \times \cdots \times P'$. Then for measurable rectangles $A_1 \times A_2 \times \cdots \times A_n \in \mathscr{S}^n$ we have

$$P_{k_1,k_2,\ldots,k_n}(A_1 \times A_2 \times \cdots \times A_n) = \prod_{i=1}^{n} P'(A_i). \tag{0.21}$$

It is easy to check that these measures are consistent (note that the consistency condition clearly holds for measurable rectangles, and then apply the π-λ theorem), and thus we may apply the Kolmogorov extension theorem and obtain a unique probability measure P on $(S^{\mathbb{Z}}, \mathscr{S}^{\mathbb{Z}})$, the space of all doubly infinite sequences consisting of zeros and ones. Let $X_t(\omega) = \omega_t$. If $k_1 < k_2 < \ldots < k_n \in \mathbb{Z}$ and $x_i \in \{0, 1\}$ for all $i \leq n$, then

$$P(X_{k_1} = x_1, X_{k_2} = x_2, \ldots, X_{k_n} = x_n) = \prod_{i=1}^{n} P'(\{x_i\}) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

If a process $(X_k)_{k\in\mathbb{N}}$ is stationary, we may use the Kolmogorov extension theorem to obtain a process $(X'_k)_{k\in\mathbb{Z}}$ such that $(X_k)_{k\in\mathbb{N}}$ and $(X'_k)_{k\in\mathbb{N}}$ are identically distributed.

Theorem 0.22. *Let $(X_k)_{k\in\mathbb{N}}$ be a stationary stochastic process with state space $(S, \mathscr{S})$ such that S is a complete, separable metric space and $\mathscr{S} = \mathscr{B}(S)$. Then there exists a probability space $(S^{\mathbb{Z}}, \mathscr{S}^{\mathbb{Z}}, Q)$ and a stochastic process $(X'_k)_{k\in\mathbb{Z}}$ defined on $(S^{\mathbb{Z}}, \mathscr{S}^{\mathbb{Z}}, Q)$ such that $(X_k)_{k\in\mathbb{N}}$ and $(X'_k)_{k\in\mathbb{N}}$ are identically distributed, that is, $P_X(A) = P_{X'}(A)$ for all $A \in \mathscr{S}^{\infty}$.*

Proof. Let $k_1 < \ldots < k_n \in \mathbb{Z}$, and let

$$P_{k_1,k_2,\ldots,k_n}(B) = P[(X_1, X_{k_2-k_1+1}, \ldots, X_{k_n-k_1+1}) \in B].$$

Since marginal distributions of a stochastic process always satisfy the consistency condition, the probability measures $P_{k_1,k_2,\ldots,k_n}$ are consistent, and thus an application of the Kolmogorov extension theorem yields the desired result (recall that the distribution of a stochastic process is uniquely determined by its marginal distributions). □

0.4 Uniform integrability

Suppose that X is an integrable real-valued random variable on a probability space $(\Omega, \mathscr{F}, P)$. Then $|X|I_{\{|X|>\alpha\}}$ is dominated by $|X|$, and the dominated converge theorem

implies that

$$\lim_{\alpha\to\infty}\int_{\{|X|>\alpha\}}|X|\,\mathrm{d}P=0. \tag{0.23}$$

Definition 0.24. We say that a sequence $(X_n)_{n\in\mathbb{N}}$ of real-valued random variables is *uniformly integrable* if (0.23) holds uniformly in n, that is,

$$\lim_{\alpha\to\infty}\sup_n\int_{\{|X_n|>\alpha\}}|X_n|\,\mathrm{d}P=0. \tag{0.25}$$

Uniform integrability implies that each X_n is integrable. To see this, let α be so large that the supremum in (0.25) is less than 1. Then

$$\int_\Omega|X_n|\,\mathrm{d}P=\int_{\{|X_n|>\alpha\}}|X_n|\,\mathrm{d}P+\int_{\{|X_n|\le\alpha\}}|X_n|\,\mathrm{d}P\le 1+\alpha<\infty. \tag{0.26}$$

Anyone who is familiar with the monotone and dominated convergence theorems knows that it is often very convenient if the order of taking a limit and integration can be reversed. Uniform integrability allows us to do it:

Theorem 0.27. *Suppose that* $\lim_{n\to\infty}X_n=X$ *almost surely. Then,*

(i) If the functions X_n *are uniformly integrable, then* X *is integrable and* $\lim_{n\to\infty}\int X_n\,\mathrm{d}P=\int X\,\mathrm{d}P$.

(ii) If X *and the* X_n *are nonnegative and integrable, then* $\lim_{n\to\infty}\int X_n\,\mathrm{d}P=\int X\,\mathrm{d}P$ *implies that the* X_n *are uniformly integrable.*

Proof. (i) By Fatou's lemma and (0.26),

$$\int_\Omega|X|\,\mathrm{d}P=\int_\Omega\liminf_n|X_n|\,\mathrm{d}P\le\liminf_n\int_\Omega|X_n|\,\mathrm{d}P\le 1+\alpha<\infty.$$

Therefore, X is integrable.

Let α be a positive real number such that $P(|X|=\alpha)=0$, and define $X_n^\alpha=X_nI_{\{|X_n|<\alpha\}}$, $X^\alpha=XI_{\{|X|<\alpha\}}$. Since $P(|X|=\alpha)=0$, we have $\lim_{n\to\infty}X_n^\alpha=X^\alpha$ with probability 1. And since the $|X_n^\alpha|$ are uniformly bounded by α, the dominated convergence theorem implies that

$$\lim_{n\to\infty}\int_\Omega X_n^\alpha\,\mathrm{d}P=\int_\Omega X^\alpha\,\mathrm{d}P. \tag{0.28}$$

Since

$$\left|\int_\Omega X_n\,\mathrm{d}P - \int_\Omega X\,\mathrm{d}P\right| = \left|\int_{\{|X_n|\geq\alpha\}} X_n\,\mathrm{d}P + \int_\Omega X_n^\alpha\,\mathrm{d}P - \int_{\{|X|\geq\alpha\}} X\,\mathrm{d}P - \int_\Omega X^\alpha\,\mathrm{d}P\right|$$
$$\leq \left|\int_\Omega X_n^\alpha\,\mathrm{d}P - \int_\Omega X^\alpha\,\mathrm{d}P\right| + \int_{\{|X_n|\geq\alpha\}} |X_n|\,\mathrm{d}P + \int_{\{|X|\geq\alpha\}} |X|\,\mathrm{d}P,$$

it follows from (0.28) that

$$\limsup_{n\to\infty}\left|\int_\Omega X_n\,\mathrm{d}P - \int_\Omega X\,\mathrm{d}P\right| \leq \sup_{n\geq 1}\int_{\{|X_n|\geq\alpha\}} |X_n|\,\mathrm{d}P + \int_{\{|X|\geq\alpha\}} |X|\,\mathrm{d}P. \tag{0.29}$$

Now pick a sequence $(\alpha_k)_{k\in\mathbb{N}}$ such that $\alpha_k\to\infty$ and $P(|X|=\alpha_k)=0$ for all k. Such a sequence exists because $P(|X|=x)$ can be positive for at most countably many x. Then $\sup_{n\geq1}\int_{\{|X_n|\geq\alpha_k\}}|X_n|\,\mathrm{d}P\to 0$ as $k\to\infty$ by uniform integrability, and because X is integrable, $\int_{\{|X|\geq\alpha_k\}}|X|\,\mathrm{d}P$ converges to zero as well. Hence, (i) follows from (0.29).

To prove the second claim, suppose that X and the X_n are nonnegative and integrable, and $\lim_{n\to\infty}\int X_n\,\mathrm{d}P = \int X\,\mathrm{d}P$ holds. If $P(X=\alpha)=0$, then (0.28) holds again, and

$$\lim_{n\to\infty}\int_{\{X_n\geq\alpha\}} X_n\,\mathrm{d}P = \lim_{n\to\infty}\left(\int_\Omega X_n\,\mathrm{d}P - \int_\Omega X_n^\alpha\,\mathrm{d}P\right) = \int_\Omega X\,\mathrm{d}P - \int_\Omega X^\alpha\,\mathrm{d}P$$
$$= \int_{\{X\geq\alpha\}} X\,\mathrm{d}P.$$

Since X is integrable, for each $\epsilon>0$ there exists an α such that $\int_{\{X\geq\alpha\}}X\,\mathrm{d}P$ is less than ϵ and $P(X=\alpha)=0$. This implies that for some $n_0\in\mathbb{N}$, the integrals $\int_{\{X_n\geq\alpha\}}X_n\,\mathrm{d}P$ are less than ϵ for all $n\geq n_0$. Since the individual X_n are integrable, we may increase α so that all the integrals are smaller than ϵ. Therefore, the X_n are uniformly integrable. □

Corollary 0.30. *If X and the X_n are integrable, and if $X_n\to X$ with probability 1, then $X_n\to X$ in L^1 if and only if the X_n are uniformly integrable.*

Proof. Suppose that the X_n are uniformly integrable. Then the differences $|X-X_n|$ are also uniformly integrable and since they converge to 0 almost surely by our hypothesis, the theorem implies that $\lim_{n\to\infty}\int_\Omega|X-X_n|\,\mathrm{d}P=0$, that is, $X_n\to X$ in L^1.

Conversely, suppose that $X_n\to X$ in L^1. Then since $||X|-|X_n||\leq|X-X_n|$, we have $\lim_{n\to\infty}\int_\Omega|X_n|\,\mathrm{d}P=\int_\Omega|X|\,\mathrm{d}P$. But then statement (ii) of the theorem implies that the $|X_n|$ are uniformly integrable. Equivalently, the X_n are then uniformly integrable. □

0.5 Conditional expectation and probability

If μ and ν are two probability measures defined on a σ-algebra $\mathscr{F}$ and $\mu(A) = 0$ implies $\nu(A) = 0$, then we say that ν is *absolutely continuous* with respect to μ, and we write $\nu << \mu$.

Lemma 0.31. *Suppose that $\nu << \mu$, and let $\epsilon > 0$ be arbitrary. Then there exists a positive real number δ such that $\nu(A) < \epsilon$ for all A such that $\mu(A) < \delta$.*

Proof. Suppose that the claim is not true. Then there exists an $\epsilon > 0$ and sets $A_1, A_2, \ldots$ such that $\mu(A_n) < \frac{1}{n^2}$ and $\nu(A_n) \geq \epsilon$ for all n. Now the Borel-Cantelli lemma implies that $\mu\left(\bigcap_{m\geq 1}\bigcup_{n\geq m} A_n\right) = 0$, but

$$\nu\left(\bigcap_{m\geq 1}\bigcup_{n\geq m} A_n\right) = \liminf_{m\to\infty} \nu\left(\bigcup_{n\geq m} A_n\right) \geq \epsilon,$$

contradicting the absolute continuity. □

Corollary 0.32. *If Y is integrable, then for each $\epsilon > 0$ there exists a $\delta > 0$ such that $\left|\int_A Y \,\mathrm{d}P\right| < \epsilon$ for all A such that $P(A) < \delta$.*

Proof. Put $\lambda(A) = \int_A |Y| \,\mathrm{d}P$. Then $\lambda << P$. Let $\epsilon > 0$ be given. Then by the theorem there exists a $\delta > 0$ such that

$$\left|\int_A Y \,\mathrm{d}P\right| \leq \lambda(A) < \epsilon$$

for all A such that $P(A) < \delta$. □

The following famous theorem states that if $\nu << \mu$, then ν can be represented as an integral with respect to μ:

Theorem 0.33. *(Radon-Nikodym Theorem) Suppose that μ and ν are probability measures defined on $(\Omega, \mathscr{F})$. If $\nu << \mu$, then there exists a μ-measurable function g such that*

$$\nu(A) = \int_A g \,\mathrm{d}\mu \text{ for all } A \in \mathscr{F}.$$

Moreover, if h is any other function with this property, then $h = g$ almost surely with respect to μ.

Proof. [1, p. 64] □

The function g is called the Radon-Nikodym derivative of ν with respect to μ. We denote

$$g = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}.$$

Radon-Nikodym derivatives have the following property:

Lemma 0.34. *(Chain rule) If $\mu_1 << \mu_2$ and $\mu_2 << \mu_3$, then*

$$\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_3} = \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_3}$$

with μ_3-probability 1.

Proof. [8, p. 241] □

Let Y be an integrable random variable defined on $(\Omega, \mathscr{F}, P)$. Suppose that $\mathscr{G}$ is a sub-σ-algebra of $\mathscr{F}$, and define a probability measure λ on $\mathscr{G}$ by $\lambda(A) = \int_A Y \,\mathrm{d}P$, $A \in \mathscr{G}$. Since $\lambda << P$, we can define the *conditional expectation of Y given $\mathscr{G}$*, denoted by $E[Y|\mathscr{G}]$, as the Radon-Nikodym derivative of λ with respect to P. In others words, we obtain a unique (up to P-measure 1) random variable $E[Y|\mathscr{G}]$ with the following properties:

(1) $E[Y|\mathscr{G}]$ is $\mathscr{G}$-measurable;
(2) $\int_A Y \,\mathrm{d}P = \int_A E[Y|\mathscr{G}] \,\mathrm{d}P$ for all $A \in \mathscr{G}$.

Remark 0.35. If $\mathscr{G} = \sigma(X)$ for some random variable X, then it is customary to write $E[Y|X]$ instead of $E[Y|\sigma(X)]$. We adapt this convention.
If $X : (\Omega, \mathscr{F}) \to (\Omega', \mathscr{F}')$ is measurable $\mathscr{F}/\mathscr{F}'$, and $Y \in L^1(\Omega)$, then the Radon-Nikodym theorem also implies that there exists a unique (up to P_X-measure 1) function $g : (\Omega', \mathscr{F}') \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ such that

$$\int_{\{X \in A\}} Y(\omega)\, P(\mathrm{d}\omega) = \int_A g(x)\, P_X(\mathrm{d}x) \text{ for all } A \in \mathscr{F}'. \tag{0.36}$$

We denote the function $g(x)$ by $E[Y|X = x]$. It is also worth noting that $E[Y|X] = g(X)$. To see this, let $A \in \sigma(X)$. Then A is of form $\{X \in A'\}$ for some $A' \in \mathscr{F}'$ and

$$\begin{aligned}\int_A Y \,\mathrm{d}P &= \int_{\{X \in A'\}} Y \,\mathrm{d}P = \int_{A'} g(x)\, P_X(\mathrm{d}x) = \int_{\Omega'} I_{A'}(x) g(x)\, P_X(\mathrm{d}x) \\ &= \int_\Omega I_{A'}[X(\omega)] g[X(\omega)]\, P(\mathrm{d}\omega) = \int_A g(X) \,\mathrm{d}P.\end{aligned}$$

For any event $A \in \mathscr{F}$, the conditional probabilities $P(A|\mathscr{G})$ and $P(A|X = x)$ are defined by $E[I_A|\mathscr{G}]$ and $E[I_A|X = x]$, respectively. If $P(X = x) > 0$, then $P(A|X = x)$ agrees with the elementary definition $P(A|B) = P(A \cap B)/P(B)$ of conditional probability. For details, see [1, p. 201].

Example 0.37. Suppose that X and Y are random variables defined on $(\Omega, \mathscr{F}, P)$, taking values in arbitrary measurable spaces. Suppose further that the random variable $Z : (\Omega, \mathscr{F}) \to (\Omega', \mathscr{F}')$ is measurable $\mathscr{F}/\mathscr{F}'$. By the elementary definition of conditional probability, we have

$$P(X = x, Y = y \mid Z = z) = P(Y = y \mid X = x, Z = z)P(X = x \mid Z = z) \tag{0.38}$$

if $P(X = x, Z = z) > 0$. We will now use the chain rule of Radon-Nikodym derivatives to show that this holds even if $P(X = x, Z = z) = 0$. Define $\mu_3 = P_Z$ and $\mu_1(A) = \int_{\{Z \in A\}} I_{\{X=x, Y=y\}} \, \mathrm{d}P$, $A \in \mathscr{F}'$. Then

$$\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_3}(z) = E\left[I_{\{X=x,Y=y\}} \mid Z = z\right] = P(X = x, Y = y \mid Z = z).$$

If probability measure μ_2 is defined on $\mathscr{F}'$ by $\mu_2(A) = \int_{\{Z \in A\}} I_{\{X=x\}} \, \mathrm{d}P$, then $\mu_2 << \mu_3$, and so the chain rule implies that

$$P(X = x, Y = y \mid Z = z) = \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}(z)\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_3}(z).$$

Clearly $\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_3}(z) = E\left[I_{\{X=x\}} \mid Z = z\right] = P(X = x \mid Z = z)$. Note also that

$$\begin{aligned}
\mu_1(A) &= \int_{\{Z \in A\}} I_{\{X=x,Y=y\}} \, \mathrm{d}P = \int_{\{Z \in A, X=x\}} I_{\{Y=y\}} \, \mathrm{d}P \\
&= \int_{A \times \{x\}} E\left[I_{\{Y=y\}} \mid X = x', Z = z'\right] \, P_{Z,X}(\mathrm{d}z', \mathrm{d}x') \\
&= \int_{A \times \{x\}} E\left[I_{\{Y=y\}} \mid X = x, Z = z'\right] \, P_{Z,X}(\mathrm{d}z', \mathrm{d}x').
\end{aligned}$$

Starting from indicator functions, it is easy to check that

$$\int_{A \times \{x\}} f(z') \, P_{Z,X}(\mathrm{d}z', \mathrm{d}x') = \int_A f(z) \, \mu_2(\mathrm{d}z)$$

holds for all integrable $\mathscr{F}'$-measurable functions f. Therefore,

$$\mu_1(A) = \int_A E\left[I_{\{Y=y\}} \mid X = x, Z = z\right] \, \mu_2(\mathrm{d}z),$$

and thus $\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}(z) = E\left[I_{\{Y=y\}} \mid X = x, Z = z\right] = P\left[Y = y \mid X = x, Z = z\right]$. This proves that (0.38) holds almost surely with respect to $\mu_3 = P_Z$.

In the next theorem we list some of the most important properties of conditional expectation. Note that since conditional expectation is not unique, the equalities and inequalities hold only almost surely.

Theorem 0.39. *Let $Y \in L^1(\Omega)$, and let $\mathscr{G}$ and $\mathscr{H}$ be sub-σ-algebras of $\mathcal{F}$. Then*

(1) (Linearity) $E[aX + bY|\mathscr{G}] = aE[X|\mathscr{G}] + bE[Y|\mathscr{G}]$;
(2) (Law of iterated expectation) $E[E[Y|\mathscr{G}]] = E[Y]$;
(3) (Tower property) If $\mathscr{G} \subset \mathscr{H}$, then $E[E[Y|\mathscr{G}] \mid \mathscr{H}] = E[E[Y|\mathscr{H}] \mid \mathscr{G}] = E[Y \mid \mathscr{G}]$;
(4) ("Taking out what is known") If X is $\mathscr{G}$-measurable and $XY \in L^1(\Omega)$, then $E[XY|\mathscr{G}] = XE[Y|\mathscr{G}]$;
(5) (Role of independence) If Y is independent of $\mathscr{G}$, then $E[Y|\mathscr{G}] = E[Y]$;
(6) $|E[Y|\mathscr{G}]| \leq E[\,|Y|\,|\mathscr{G}]$;
(7) If $X \leq Y$ almost surely, then $E[X|\mathscr{G}] \leq E[Y|\mathscr{G}]$ almost surely.

Proof. See [1, p. 220]. □

0.6 Martingales

A certain amount of martingale convergence theory is involved in the proof the Shannon-McMillan-Breiman theorem.

Definition 0.40. Let $\mathscr{F}$ be a σ-algebra. A collection $(\mathscr{F}_n)_{n\in\mathbb{N}}$ of sub-σ-algebras of $\mathscr{F}$ is called a *filtration* if $\mathscr{F}_i \subset \mathscr{F}_j$ for all $i \leq j$. A stochastic process $X = (X_n)_{n\in\mathbb{N}}$ is *adapted* to the filtration if for every $n \in \mathbb{N}$, X_n is $\mathscr{F}_n$-measurable. The *natural filtration* $(\mathscr{F}_n^X)_{n\in\mathbb{N}}$ of the process X is defined by $\mathscr{F}_n^X = \sigma(X_m : m \leq n)$.

Note that every stochastic process is clearly adapted to its natural filtration.

Definition 0.41. Let $X = (X_n)_{n\in\mathbb{N}}$ be a stochastic process defined on $(\Omega, \mathscr{F}, P)$. Suppose that X is adapted to the filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$. Suppose further that each $X_n, n \in \mathbb{N}$, is integrable. If $E(X_{n+1}|\mathscr{F}_n) = X_n$ for all $n \geq 1$, then we say that X is a *martingale relative to the $\mathscr{F}_n$*. If $E(X_{n+1}|\mathscr{F}_n) \geq X_n$, then X is called a *submartingale*, and if $E(X_{n+1}|\mathscr{F}_n) \leq X_n$, then X is called a *supermartingale*.

The properties of conditional expectation imply that for martingales we have $E[X_n] = E[X_0]$ for all $n \in \mathbb{N}$. For supermartingales and submartingales we have $E[X_n] \leq E[X_0]$ and $E[X_n] \geq E[X_0]$, respectively.

The martingale constructed in the following example will play an important role in the proof of the Shannon-McMillan-Breiman theorem.

Example 0.42. Let $\mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots$ be an increasing sequence of sub-σ-algebras of $\mathscr{F}$, and let X be an integrable random variable. Then the process $Y_n = E[X|\mathscr{F}_n]$ is a martingale relative to the $\mathscr{F}_n$: the Y_n are integrable since $E[Y_n] = E[E[X|\mathscr{F}_n]] = E[X] < \infty$, and

$$E[Y_{n+1}|\mathscr{F}_n] = E[E[X|\mathscr{F}_{n+1}]|\mathscr{F}_n] = E[X|\mathscr{F}_n] = Y_n$$

by the "tower property" of conditional expectation.

In particular, let $(X_k)_{k\in\mathbb{Z}}$ be a stochastic process, and let $\mathscr{F}_n = \sigma(X_{-1}, X_{-2}, \ldots, X_{-n})$. Then the $\mathscr{F}_n$ form an increasing sequence of σ-algebras, and if $Z = I_{\{X_0=x_0\}}$, then the sequence $E[Z|\mathscr{F}_n] = P(X_0 = x_0|X_{-1}, X_{-2}, \ldots, X_{-n}), n \geq 1$, is a martingale.

Convergence of martingales

We will next study the conditions under which martingales converge.

Definition 0.43. A stochastic process $X = (X_n)_{n\in\mathbb{N}}$ is called *predictable* with respect to the filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$ if X_n is $\mathscr{F}_{n-1}$-measurable for every $n \geq 2$ and X_1 is constant.

Definition 0.44. Let M and X be two stochastic processes with parameter set $\mathbb{N}$. The process $X \cdot M$ is defined by $(X \cdot M)_1 = 0$ and

$$(X \cdot M)_n = \sum_{k=2}^{n} X_k(M_k - M_{k-1})$$

for $n \geq 2$. The process $(X \cdot M)$ is called the *discrete integral of X with respect to M.*

Lemma 0.45. *If M is supermartingale and X is a bounded, nonnegative predictable process, then $X \cdot M$ is a supermartingale as well. If M is a martingale and X is a bounded, predictable process, then $X \cdot M$ is a martingale.*

Proof. Put $Y = X \cdot M$. It is clear that Y_n is $\mathscr{F}_n$-measurable, and since $|X_n| < K < \infty$ for some $K \in \mathbb{R}$ and for all $n \in \mathbb{N}$, we have $E[|Y_n|] \leq 2K \sum_{k=1}^{n} E[|M_k|] < \infty$. If M is a supermartingale and X is a nonnegative predictable process, then

$$\begin{aligned} E[Y_n|\mathscr{F}_{n-1}] &= E[Y_{n-1} + X_n(M_n - M_{n-1})|\mathscr{F}_{n-1}] = Y_{n-1} + X_n E[M_n - M_{n-1}|\mathscr{F}_{n-1}] \\ &= Y_{n-1} + X_n(E[M_n|\mathscr{F}_{n-1}] - M_{n-1}) \leq Y_{n-1} + X_n(M_{n-1} - M_{n-1}) = Y_{n-1}. \end{aligned}$$

If M is a martingale, the inequality is an equality regardless of the sign of X_n, and thus in this case Y is a martingale. □

Suppose that M is a supermartingale and consider a closed interval $[a, b] \subset \mathbb{R}$, $a < b$. The number of *upcrossings* of $[a, b]$ that the process M makes up to time n, denoted by $U_n[a, b]$, is the number of times the process moves from a level below a to a level above b. To be precise:

Definition 0.46. Let M be a supermartingale, and let $a < b \in \mathbb{R}$. The number $U_n[a,b](\omega)$ is the largest $k \in \mathbb{N}_0$ such that there exist $0 < s_1 < t_1 < s_2 < t_2 < \cdots < s_k < t_k \leq n$ with $X_{s_i}(\omega) < a$ and $X_{t_i}(\omega) > b$.

Lemma 0.47. *(Doob's upcrossings lemma) Let M be a supermartingale. Then for all $a < b \in \mathbb{R}$ and $n \in \mathbb{N}$ we have*

$$(b-a)E[U_n[a,b]] \leq E[(M_n - a)^-].$$

Proof. Define a bounded, nonnegative predictable process $X = (X_k)_{k\in\mathbb{N}}$ by $X_1 = 0, X_2 = I_{\{M_1<a\}}$ and

$$X_n = I_{\{X_{n-1}=1\}} I_{\{M_{n-1}\leq b\}} + I_{\{X_{n-1}=0\}} I_{\{M_{n-1}<a\}}$$

for $n \geq 3$. Let $Y = X \cdot M$. By the previous lemma, Y is a supermartingale. Note that the process X is 0 until M drops below the level a, and then it is 1 until M gets above b, and so on. Therefore, every completed upcrossing increases Y by at least $b-a$. If the last upcrossing has not been completed at time n, this can cause Y to decrease by at most $(M_n - a)^-$. Hence,

$$Y_n \geq (b-a)U_n[a,b] - (M_n - a)^-.$$

Since Y_n is a supermartingale, we have $E[Y_n] \leq E[Y_0] = 0$, which finally implies that

$$0 \geq E[Y_n] \geq (b-a)E[U_n[a,b]] - E[(M_n - a)^-].$$

□

Here is our first martingale convergence theorem, due to J.L. Doob:

Theorem 0.48. *(Doob's martingale convergence theorem) Suppose that $M = (M_n)_{n\in\mathbb{N}}$ is a supermartingale and bounded in L^1. Then M_n converges almost surely to a limit M_∞ as $n \to \infty$, and M_∞ is integrable.*

Proof. Let $\Lambda_{a,b} = \{\omega : \liminf_{n\to\infty} M_n < a < b < \limsup_{n\to\infty} M_n\}$, and let Λ be the set on which M does not converge. Then, clearly

$$\Lambda = \{\omega : \liminf_{n\to\infty} M_n < \limsup_{n\to\infty} M_n\} = \bigcup_{\substack{a<b \\ a,b\in\mathbb{Q}}} \{\omega : \liminf_{n\to\infty} M_n < a < b < \limsup_{n\to\infty} M_n\}$$

$$= \bigcup_{\substack{a<b \\ a,b\in\mathbb{Q}}} \Lambda_{a,b}.$$

Define $U_\infty[a,b] = \lim_{n\to\infty} U_n[a,b]$. Then $\Lambda_{a,b} \subset \{\omega : U_\infty[a,b](\omega) = \infty\}$. Since the M_n are bounded in L^1, we have $\sup_n E[|M_n|] = K < \infty$ for some $K \in \mathbb{R}$. The monotone convergence theorem and Doob's upcrossings lemma then imply that

$$E[U_\infty[a,b]] = E[\lim_{n\to\infty} U_n[a,b]] = \lim_{n\to\infty} E[U_n[a,b]] \leq \frac{K+|a|}{b-a} < \infty.$$

Therefore, $U_\infty[a,b]$ is finite almost surely which implies that $P(\Lambda_{a,b}) = 0$. We also have $P(\Lambda) = 0$ since Λ is a countable union of sets of P-measure 0. Therefore, $\lim_{n\to\infty} M_n = M_\infty$ exists almost surely. Finally, by Fatou's lemma,

$$E[|M_\infty|] = E[\liminf_{n\to\infty} |M_n|] \leq \liminf_{n\to\infty} E[|M_n|] \leq \sup_{n\geq 1} E[|M_n|] < \infty.$$

□

If we also assume uniform integrability, then M_n not only converges almost surely but also in L^1:

Theorem 0.49. *Suppose that $M = (M_n)_{n\in\mathbb{N}}$ is a supermartingale and bounded in L^1. Then $M_n \to M_\infty$ in L^1 if and only if the M_n are uniformly integrable. In this case,*

$$E[M_\infty|\mathscr{F}_n] \leq M_n$$

almost surely for all $n \in \mathbb{N}$, and if M is a martingale, equality holds.

Proof. By the previous theorem, M_n converges to M_∞ almost surely and M_∞ is integrable. Then, by Corollary 0.30, $M_n \to M_\infty$ in L^1 if and only if the M_n are uniformly integrable. As for the second claim, suppose that $M_n \to M_\infty$ in L^1, and let $n \in \mathbb{N}$. Since M is a supermartingale, we have

$$E[M_m I_A] = E[E[M_m I_A|\mathscr{F}_n]] = E[I_A E[M_m|\mathscr{F}_n]] \leq E[M_n I_A] \tag{0.50}$$

for all $A \in \mathscr{F}_n$ and $m > n$. But

$$|E[M_m I_A] - E[M_\infty I_A]| \leq E[|M_m - M_\infty|],$$

and $E[|M_m - M_\infty|]$ converges to zero as $m \to \infty$ by the L^1 convergence. Therefore, $E[M_m I_A]$ converges to $E[M_\infty I_A]$ as $m \to \infty$ and by (0.50), the limit must be less than or equal to $E[M_n I_A]$. Now

$$\int_A E[M_\infty|\mathscr{F}_n]\,\mathrm{d}P = \int_A M_\infty\,\mathrm{d}P \leq \int_A M_n\,\mathrm{d}P$$

for all $A \in \mathscr{F}_n$, and thus $E[M_\infty|\mathscr{F}_n] \leq M_n$ almost surely. Finally, equality holds in (0.50) if M is a martingale, and in this case, $E[M_\infty|\mathscr{F}_n] = M_n$ almost surely. □

Lemma 0.51. *Let Y be an integrable random variable on $(\Omega, \mathscr{F}, P)$, and let $\mathscr{F}_n, n \in \mathbb{N}$, be a collection of sub-σ-algebras of $\mathscr{F}$. Then the random variables $E[Y|\mathscr{F}_n], n \in \mathbb{N}$, are uniformly integrable.*

Proof. Let $c > 0$. First we observe that

$$\int_{\{|E[Y|\mathscr{F}_n]| \geq \alpha\}} |E[Y|\mathscr{F}_n]| \, \mathrm{d}P \leq \int_{\{|E[Y|\mathscr{F}_n]| \geq \alpha\}} E[\,|Y|\,|\mathscr{F}_n] \, \mathrm{d}P = \int_{\{|E[Y|\mathscr{F}_n]| \geq \alpha\}} |Y| \, \mathrm{d}P. \tag{0.52}$$

Then we apply the Chebyshev inequality:

$$P(|E[Y|\mathscr{F}_n]| \geq \alpha) \leq \frac{E[|E[Y|\mathscr{F}_n]|]}{\alpha} \leq \frac{E[E[\,|Y|\,|\mathscr{F}_n]]}{\alpha} = \frac{E[|Y|]}{\alpha}.$$

The upper bound does not depend on n, and it can be made arbitrarily small by increasing α. Since Y is integrable, uniform integrability follows now from (0.52) and Lemma 0.32. □

We may now prove Levy's martingale convergence theorem, which will be used later in the proof of Shannon-McMillan-Breiman theorem.

Theorem 0.53. *(Levy's martingale convergence theorem) Let $(\mathscr{F}_n)_{n\in\mathbb{N}}$ be an increasing sequence of sub-σ-algebras of $\mathscr{F}$, and let $\mathscr{F}_\infty$ be the σ-algebra generated by $\bigcup_{n=1}^\infty \mathscr{F}_n$. If Y is integrable, then $E[Y|\mathscr{F}_n] \to E[Y|\mathscr{F}_\infty]$ as $n \to \infty$ almost surely and in L^1.*

Proof. By Example 0.42 and the previous lemma, the sequence $(E[Y|\mathscr{F}_n])_{n\in\mathbb{N}}$ is an uniformly integrable martingale. Since $E[|E[Y|\mathscr{F}_n]|] \leq E[E[\,|Y|\,|\mathscr{F}_n]] = E[|Y|] < \infty$, the sequence is bounded in L^1, and thus Doob's martingale convergence theorem implies that it converges almost surely to an integrable random variable X_∞. Convergence in L^1 follows from Theorem 0.49.

It remains to be shown that $X_\infty = E[Y|\mathscr{F}_\infty]$. First we check that the generating set of $\mathscr{F}_\infty$, $\bigcup_{n=1}^\infty \mathscr{F}_n$, is a π-system. If $F_1, F_2 \in \bigcup_{n=1}^\infty \mathscr{F}_n$, then $F_1 \in \mathscr{F}_{n_1}$ and $F_2 \in \mathscr{F}_{n_2}$ for some n_1 and n_2. Suppose that $n_1 < n_2$. Then $F_1 \in \mathscr{F}_{n_1} \subset \mathscr{F}_{n_2}$, and since $\mathscr{F}_{n_2}$ is a σ-algebra, $F_1 \cap F_2 \in \mathscr{F}_{n_2} \subset \bigcup_{n=1}^\infty \mathscr{F}_n$. Similarly, if $n_2 \leq n_1$, then $F_1 \cap F_2 \in \mathscr{F}_{n_1} \subset \bigcup_{n=1}^\infty \mathscr{F}_n$. Thus $\bigcup_{n=1}^\infty \mathscr{F}_n$ is a π-system.

Suppose that $A \in \bigcup_{n=1}^\infty \mathscr{F}_n$. Then $A \in \mathscr{F}_n$ for some $n \in \mathbb{N}$, and the L^1 convergence implies that

$$\int_A Y \, \mathrm{d}P = \lim_{m\to\infty} \int_A E[Y|\mathscr{F}_m] \, \mathrm{d}P = \int_A X_\infty \, \mathrm{d}P \tag{0.54}$$

Thus $\int_A Y \, \mathrm{d}P = \int_A X_\infty \, \mathrm{d}P$ holds for all $A \in \bigcup_{n=1}^\infty \mathscr{F}_n$. If we can show that the class of sets $\mathscr{C}$ for which it holds is a λ-system, then the π-λ theorem implies that it holds for all $A \in \mathscr{F}_\infty$. But this is easy:

(1) Since Ω belongs to $\bigcup_{n=1}^{\infty} \mathscr{F}_n$, $\int_\Omega Y \,\mathrm{d}P = \int_\Omega X_\infty \,\mathrm{d}P$.
(2) Suppose that $A \in \mathscr{C}$. Then

$$\int_{A^c} Y \,\mathrm{d}P = \int_\Omega Y \,\mathrm{d}P - \int_A Y \,\mathrm{d}P = \int_\Omega X_\infty \,\mathrm{d}P - \int_A X_\infty \,\mathrm{d}P = \int_{A^c} X_\infty \,\mathrm{d}P,$$

which implies that $A^c \in \mathscr{C}$.
(3) Suppose that $A_1, A_2, \ldots$ are disjoint $\mathscr{C}$-sets. Then

$$\int_{\bigcup_{n=1}^{\infty} A_n} Y \,\mathrm{d}P = \sum_{n=1}^{\infty} \int_{A_n} Y \,\mathrm{d}P = \sum_{n=1}^{\infty} \int_{A_n} X_\infty \,\mathrm{d}P = \int_{\bigcup_{n=1}^{\infty} A_n} X_\infty \,\mathrm{d}P,$$

which implies that $\bigcup_{n=1}^{\infty} A_n$ belongs to $\mathscr{C}$.

Hence, $\mathscr{C}$ is a λ-system, and thus $\int_A Y \,\mathrm{d}P = \int_A X_\infty \,\mathrm{d}P$ holds for all $A \in \sigma\left(\bigcup_{n=1}^{\infty} \mathscr{F}_n\right) = \mathscr{F}_\infty$. Since $E[Y|\mathscr{F}_n]$ is $\mathscr{F}_n \subset \mathscr{F}_\infty$-measurable, X_∞ as a limit of $\mathscr{F}_\infty$-measurable functions is also $\mathscr{F}_\infty$-measurable. Now $X_\infty = E[Y|\mathscr{F}_\infty]$ by the definition of conditional expectation. □

Chapter 1

Ergodic Theory

1.1 Introduction

Ergodic theory could be described as the study of the long term average behaviour of systems evolving in time. Consider the following examples.

Suppose that $X_1, X_2, X_3, \ldots$ are independent, identically distributed random variables with finite mean m. Then the Strong Law of Large Numbers states that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = m$$

almost surely.

Consider an aperiodic, irreducible Markov chain $X_1, X_2, X_3, \ldots$ with stationary distribution π and finite state space S. Then, if $j \in S$,

$$\lim_{n\to\infty} \frac{\#\{1 \le k \le n : X_k = j\}}{n} = \pi(j)$$

almost surely.

Both of these results are almost immediate consequences of Birkhoff's ergodic theorem which will be proved in this chapter.

Measure-preserving transformations

First we define a basic notion of ergodic theory: the measure-preserving transformation.

Definition 1.1. Let $(\Omega, \mathscr{F}, P)$ be a probability space. A function $T : \Omega \to \Omega$ is *measurable transformation* on Ω if it is measurable $\mathscr{F}/\mathscr{F}$, that is $T^{-1}A \in \mathscr{F}$ for all $A \in \mathscr{F}$. If T is one-to-one and onto, we say that T is *invertible*. The transformation T is said to be *measure-preserving* if we have $P(T^{-1}A) = P(A)$ for all $A \in \mathscr{F}$.

Remark 1.2. If $(\Omega, \mathscr{F}, P)$ is a probability space and T is a measure-preserving transformation, then in ergodic theory the 4-tuple $(\Omega, \mathscr{F}, P, T)$ is often called a *dynamical system.* Also, the sequence $(\omega, T(\omega), T^2(\omega), \ldots)$ is called the *orbit* of ω under T.

It follows by induction that $P(T^{-k}A) = P(A)$, $k \in \mathbb{N}$, for measure-preserving transformations. In the invertible case we also have $P(T^k A) = P(A)$ for all $k \in \mathbb{N}$.

Checking whether a given transformation preserves measure, or not, can sometimes be difficult. However, the following lemma is of great help:

Lemma 1.3. *Suppose that $\mathscr{C}$ is π-system and $\mathscr{F} = \sigma(\mathscr{C})$. If T is a measurable transformation and $P(T^{-1}A) = P(A)$ for all $A \in \mathscr{C}$, then T is measure-preserving.*

Proof. Let $\mathscr{G}$ be the collection of $\mathscr{F}$-sets G for which $P(T^{-1}G) = P(G)$. We show that $\mathscr{G}$ is a λ-system. Clearly $P(T^{-1}\Omega) = P(\Omega)$. If $G \in \mathscr{G}$, then $P(T^{-1}G^c) = P[(T^{-1}G)^c] = 1 - P(T^{-1}G) = 1 - P(G) = P(G^c)$ and thus $G^c \in \mathscr{G}$. If $G_1, G_2, \ldots$ are disjoint $\mathscr{G}$-sets, then

$$P\left(T^{-1}\bigcup_{i=1}^{\infty} G_i\right) = P\left(\bigcup_{i=1}^{\infty} T^{-1}G_i\right) = \sum_{i=1}^{\infty} P(T^{-1}G_i) = \sum_{i=1}^{\infty} P(G_i) = P\left(\bigcup_{i=1}^{\infty} G_i\right).$$

Thus $\bigcup_{i=1}^{\infty} G_i \in \mathscr{G}$. We have shown that $\mathscr{G}$ is a λ-system. By hypothesis, $\mathscr{C} \in \mathscr{G}$ and therefore the π-λ theorem implies that $\sigma(\mathscr{C}) \subset \mathscr{G}$. □

Let us now consider some examples of measure-preserving transformations.

1. (Angle doubling) Let Ω be the semiclosed interval $(0, 1]$, $\mathscr{F} = \mathscr{B}((0, 1])$, and let m be the Lebesgue measure. Take $T(\omega) = 2\omega \pmod 1$:

$$T(\omega) = \begin{cases} 2\omega & \text{if } 0 < \omega \le \frac{1}{2}, \\ 2\omega - 1 & \text{if } \frac{1}{2} < \omega \le 1. \end{cases}$$

   Let $d_k(\omega)$ be the kth digit of the binary expansion of ω. Then ω has representations $\omega = 0.d_1(\omega)d_2(\omega)\cdots$ and $\omega = \sum_{k=1}^{\infty} d_k(\omega)2^{-k}$. Binary expansions are not unique; for numbers of form $\omega = \sum_{k=1}^{m-1} j_k 2^{-k} + 2^{-m}$, $m \in \mathbb{N}, j_k \in \{0, 1\}$, there are two equal expansions $0.j_1 \cdots j_{m-1}1000\cdots$ and $0.j_1 \cdots j_{m-1}0111\cdots$. Let $d_k(\omega)$ correspond to the latter *nonterminating* ones.

   The transformation T shifts the binary digits of ω to the left: $T(\omega) = 0.d_2(\omega)d_3(\omega)\cdots$. To see this, suppose first that $\omega \le \frac{1}{2}$. Then $d_1(\omega) = 0$ and

$$T(\omega) = 2 \cdot \sum_{k=1}^{\infty} d_k(\omega)2^{-k} = \sum_{k=2}^{\infty} d_k(\omega)2^{-k+1} = \sum_{k=1}^{\infty} d_{k+1}(\omega)2^{-k} = 0.d_2(\omega)d_3(\omega)\cdots.$$

If $\omega > \frac{1}{2}$, then $d_1(\omega) = 1$ and

$$T(\omega) = 2 \cdot \sum_{k=1}^{\infty} d_k(\omega)2^{-k} - 1 = \sum_{k=1}^{\infty} d_k(\omega)2^{-k+1} - 1 = \sum_{k=0}^{\infty} d_{k+1}(\omega)2^{-k} - 1$$
$$= \sum_{k=1}^{\infty} d_{k+1}(\omega)2^{-k} = 0.d_2(\omega)d_3(\omega)\cdots .$$

A *dyadic interval* is an interval of form

$$(\sum_{k=1}^{m} 2^{-j_k}, \sum_{k=1}^{m} 2^{-j_k} + 2^{-m}] = \{\omega \in (0,1] : d_1(\omega) = j_1, \ldots, d_m(\omega) = j_m\}$$

for some $m \in \mathbb{N}$ and $j_1, j_2, \ldots, j_m \in \{0,1\}$ (note that the length of the interval is 2^{-m}). Observe that since all open sets in $(0,1]$ are countable unions of dyadic intervals, $\mathscr{B}((0,1])$ is generated by the dyadic intervals. If $A = \{\omega : d_1(\omega) = a_1, \ldots, d_m(\omega) = a_m\}$ is a dyadic interval, then

$$\begin{aligned} T^{-1}A &= \{\omega : d_2(\omega) = a_1, \ldots, d_{m+1}(\omega) = a_m\} \\ &= \{\omega : d_1(\omega) = 0, d_2(\omega) = a_1, \ldots, d_{m+1}(\omega) = a_m\} \\ &\cup \{\omega : d_1(\omega) = 1, d_2(\omega) = a_1, \ldots, d_{m+1}(\omega) = a_m\}. \end{aligned}$$

Thus $T^{-1}A$ is a disjoint union of two dyadic intervals and it is definitely a Borel set. Since the dyadic intervals and the empty set form a π-system generating $\mathscr{B}((0,1])$, T is measurable by Lemma 0.11. We also conclude that

$$m(A) = 2^{-m} = 2^{-m-1} + 2^{-m-1} = m(T^{-1}A).$$

Thus T is measure-preserving by Lemma 1.3.

This transformation is called *angle doubling* since the function $f(x) = e^{2\pi ix}$ is a one-to-one mapping of $(0,1]$ onto the unit circle, and $T(\omega)$ corresponds to doubling of angle on the unit circle.

2. (Permutations) Let Ω be a finite set $\{a,b,c,d\}$ with $\mathscr{F}$ constisting of all subsets of Ω. If T is the cyclic permutation $(abcd)$ on Ω, then it is clear that T is measure-preserving if and only P assigns equal probabilities to the four points.

   If $T = (ab)(cd)$, a product of two cycles, then T is measure-preserving if and only if $P(\{a\}) = P(\{b\})$ and $P(\{c\}) = P(\{d\})$. In general, if T is any permutation of a finite set, then T can be expressed as a product of disjoint cyclic permutations. Then T is measure-preserving if and only if P assigns equal probability to each point within each cycle.

Perhaps the most important transformation is the *shift*:

Definition 1.4. Let $\Omega = S^\infty$, the space consisting of all sequences $\omega = (\omega_1, \omega_2, \ldots)$ with $\omega_k \in S$ for all $k \in \mathbb{N}$; take $\mathscr{F} = \mathscr{S}^\infty$, and let P be any probability measure on $\mathscr{S}^\infty$. If T is defined by

$$T(\omega_1, \omega_2, \ldots) = (\omega_2, \omega_3, \ldots),$$

then T is called the *one-sided shift transformation*.

The *two-sided shift transformation* is defined analogously on $S^{\mathbb{Z}}$: $(T(\omega))_k = \omega_{k+1}$, or

$$T(\ldots, \omega_{-1}, \omega_0; \omega_1, \omega_2, \ldots) = (\ldots, \omega_0, \omega_1; \omega_2, \omega_3, \ldots).$$

Since the inverse images of cylinders are cylinders (it is easy to check that $T^{-1}(\mathcal{C}(B)) = \mathcal{C}(S \times B)$), it follows by Lemma 0.11 that the shift transformations are measurable. Observe that the two-sided shift is invertible, but the one-sided shift is not.

Whether T is measure-preserving or not depends on the stationarity of the canonical stochastic process associated with S^∞ (or $S^{\mathbb{Z}}$):

Theorem 1.5. *Shift transformations are measure-preserving if and only if the stochastic process defined by $X_k(\omega) = \omega_k$ is stationary.*

Proof. Consider the one-sided case first. Suppose that the shift transformation T is measure-preserving. Let $k_1 < k_2 < \ldots < k_n \in \mathbb{N}$ and $B \in \mathscr{S}^n$. Then for all $m \in \mathbb{N}$,

$$\begin{aligned} P[(X_{k_1}, \ldots, X_{k_n}) \in B] &= P(\{\omega : (\omega_{k_1}, \ldots, \omega_{k_n}) \in B\}) \\ &= P(T^{-m}\{\omega : (\omega_{k_1}, \ldots, \omega_{k_n}) \in B\}) \\ &= P(\{\omega : (\omega_{k_{1+m}}, \ldots, \omega_{k_{n+m}}) \in B\}) \\ &= P[(X_{k_{1+m}}, \ldots, X_{k_{n+m}}) \in B]. \end{aligned}$$

Therefore, the process $(X_k)_{k\in\mathbb{N}}$ is stationary.

Conversely, suppose that the process defined by $X_k(\omega) = \omega_k$ is stationary. Let $k_1 < k_2 < \ldots < k_n \in \mathbb{N}$, and let $\mathcal{C}(B)$ be a cylinder with base B at $(k_1, k_2, \ldots, k_n)$. Then

$$\begin{aligned} P(\mathcal{C}(B)) &= P(\{\omega : (\omega_{k_1}, \ldots, \omega_{k_n}) \in B\}) = P[(X_{k_1}, \ldots, X_{k_n}) \in B] \\ &= P[(X_{k_1+1}, \ldots, X_{k_n+1}) \in B] = P(\{\omega : (\omega_{k_1+1}, \ldots, \omega_{k_n+1}) \in B\}) \\ &= P(T^{-1}\{\omega : (\omega_{k_1}, \ldots, \omega_{k_n}) \in B\}) = P(T^{-1}\mathcal{C}(B)). \end{aligned}$$

Since $\mathscr{S}^\infty$ is generated by the cylinder sets, T is measure-preserving by Lemma 1.3.

The two-sided case is proved in the same way. □

Consider the probability space $(S^T, \mathscr{S}^T, P_X)$ associated with a stochastic process $X = (X_k)_{k\in T}$ defined on $(\Omega, \mathscr{F}, P)$. By Theorem 1.5, the shift operator on S^T preserves P_X if and only if the stochastic process X is stationary. This fact will later enable us to apply the ergodic theorem for stochastic processes even if the underlying probability space is left unspecified.

Example 1.6. (Markov and Bernoulli shifts) Consider an irreducible, aperiodic Markov chain $X = (X_k)_{k\in\mathbb{N}}$ with finite state space S. If the initial distribution of X coincides with its stationary distribution, then the process is stationary, and by the previous theorem, the shift operator T on $(S^\infty, (\mathcal{P}(S))^\infty, P_X)$ is measure-preserving. Moreover, if the Markov chain is defined on the probability space $(S^\infty, (\mathcal{P}(S))^\infty, P)$ constructed in Section 0.3, then the shift operator on this space is also measure-preserving.

The sequence space of Bernoulli trials of Example 0.20 is stationary since it is a sequence consisting of independent, identically distributed trials. Therefore, the shift operators on both $(S^{\mathbb{Z}}, \mathscr{S}^{\mathbb{Z}}, P)$ and $(S^{\mathbb{Z}}, \mathscr{S}^{\mathbb{Z}}, P_X)$ are measure-preserving.

1.2 Ergodicity and mixing

It is assumed throughout this section that T is a measure-preserving transformation.

Definition 1.7. Let $A \in \mathscr{F}$. Then A is *invariant* under T if $T^{-1}A = A$. If $0 < P(A) < 1$, then A is called a *nontrivial* invariant set. If in $\mathscr{F}$ there are no nontrivial invariant sets, then T is called *ergodic*. If

$$\lim_{n\to\infty} P(A \cap T^{-n}B) = P(A)P(B) \tag{1.8}$$

for all $A, B \in \mathscr{F}$, then we say that T is *mixing*.

Mixing is stronger condition than ergodicity, as we shall now see.

Theorem 1.9. *If T is mixing, then T is ergodic.*

Proof. Suppose that B is an invariant set. Then $P(B) = P(B \cap B) = P(B \cap T^{-n}B)$ for all $n \geq 1$. Therefore,

$$P(B) = \lim_{n\to\infty} P(B \cap T^{-n}B) = P(B)P(B).$$

Thus $P(B)$ must be either 0 or 1, and we conclude that T is ergodic. □

Example 1.10. Consider again the permutation $T = (ab)(cd)$ on $\{a, b, c, d\}$. We concluded that if T is measure-preserving, then $P(\{a\}) = P(\{b\})$ and $P(\{c\}) = P(\{d\})$. If both of these probabilities are positive, then since the sets $\{a, b\}$ and $\{c, d\}$ are invariant, T is not ergodic. By the previous theorem, T can not be mixing. However, if $P(\{a, b\}) = 0$, it is easy to check that T is ergodic, but since $P(\{c\} \cap T^{-n}\{d\})$ varies between zero and $\frac{1}{2}$, T is not mixing.

If $T = (abcd)$, then T is ergodic since the only invariant sets are $\{a, b, c, d\}$ and the empty set. But T is not mixing: since T is measure-preserving, equal probabilities $\frac{1}{4}$ must be assigned to all points and thus $P(\{a\} \cap T^{-n}\{b\})$ varies between zero and $\frac{1}{4}$.

It is very convenient that it is enough to check that the mixing condition holds on a generating π-system:

Theorem 1.11. *Suppose that $\mathscr{P}$ is a π-system and $\mathscr{F} = \sigma(\mathscr{P})$. If*

$$\lim_{n\to\infty} P(A \cap T^{-n}B) = P(A)P(B)$$

for all $A, B \in \mathscr{P}$, then T is mixing.

Proof. Let $A \in \mathscr{P}$ be fixed, and let $\mathscr{C}_A$ be the class of $\mathscr{F}$-sets B for which the mixing condition (1.8) holds. We will now check that $\mathscr{C}_A$ is a λ-system. Of course,

$$\lim_{n\to\infty} P(A \cap T^{-n}\Omega) = P(A) = P(A)P(\Omega),$$

which implies that $\Omega \in \mathscr{C}_A$. Suppose then that $B \in \mathscr{C}_A$. Now we have

$$\begin{aligned} P(A \cap T^{-n}B^c) &= P(A \cap (T^{-n}B)^c) = P[(A^c \cup (T^{-n}B))^c] = 1 - P(A^c \cup T^{-n}B) \\ &= 1 - P[A^c \cup (A \cap T^{-n}B)] = 1 - P(A \cap T^{-n}B) - P(A^c). \end{aligned}$$

Letting $n \to \infty$, we obtain

$$\lim_{n\to\infty} P(A \cap T^{-n}B^c) = 1 - P(A)P(B) - P(A^c) = P(A) - P(A)P(B) = P(A)P(B^c),$$

which implies that $B^c \in \mathscr{C}_A$. Next, suppose that $B_1, B_2, \ldots$ are disjoint sets in $\mathscr{C}_A$. Then

$$\begin{aligned} \lim_{n\to\infty} P\left(A \cap T^{-n}\bigcup_{m=1}^{\infty} B_m\right) &= \lim_{n\to\infty} P\left(A \cap \bigcup_{m=1}^{\infty} T^{-n}B_m\right) = \lim_{n\to\infty} \sum_{m=1}^{\infty} P\left(A \cap T^{-n}B_m\right) \\ &= \sum_{m=1}^{\infty} \lim_{n\to\infty} P\left(A \cap T^{-n}B_m\right) = \sum_{m=1}^{\infty} P(A)P(B_m) \\ &= P(A)P\left(\bigcup_{m=1}^{\infty} B_m\right). \end{aligned}$$

The third equality is justified by the Weierstrass M-test, a well-known result in analysis ($P(A \cap T^{-n}B_m) \leq P(T^{-n}B_m) = P(B_m)$ and $\sum_{m=1}^{\infty} P(B_m)$ converges). We conclude that $\bigcup_{m=1}^{\infty} B_m \in \mathscr{C}_A$. Thus $\mathscr{C}_A$ is a λ-system and by the π-λ theorem, $\mathscr{F} = \mathscr{C}_A$.

In a similar fashion one shows that the class of $\mathscr{F}$-sets A for which the mixing condition holds for all $B \in \mathscr{F}$ is a λ-system. The π-λ theorem then implies that the mixing condition holds for all $A, B \in \mathscr{F}$, and we conclude that T is mixing. □

The easiest way to prove ergodicity is often to show that the given transformation is mixing. Let us now consider some examples.

Example 1.12. The previous theorem will now be used to show that the Markov shift is ergodic and mixing given that the Markov chain is irreducible and aperiodic, and its initial distribution coincides with its stationary distribution.

Consider the space $(S^\infty, \mathscr{S}^\infty, P)$ constructed in Section 0.3, that is, the sequence space in which $X_k(\omega) = \omega_k$, $\omega \in \Omega, k \in \mathbb{N}$, defines a Markov chain with finite state space S, initial distribution $u(i), i \in S$, and transition probabilities $p(i,j), i,j \in S$. Suppose further that the Markov chain is irreducible and aperiodic, and denote the unique stationary distribution by $\pi(i), i \in S$. We also assume that $u(i) = \pi(i)$ for all $i \in S$.

Since the state space S is finite, it is clear that all measurable rectangles in S^∞ can be written as finite disjoint unions of *thin cylinders*, that is, sets of form

$$\{\omega \in S^\infty : \omega_1 = x_1, \omega_2 = x_2, \ldots, \omega_n = x_n\},\ n \in \mathbb{N},\ x_1, x_2, \ldots, x_n \in S.$$

Therefore, if the mixing condition holds for all thin cylinders, it must hold also for all measurable rectangles. And since the measurable rectangles form a π-system generating $\mathscr{S}^\infty$, it is enough to show that the mixing condition holds for all thin cylinders.

Let T be the shift operator on S^∞, and let $A = \{\omega : \omega_1 = a_1, \ldots, \omega_{n_A} = a_{n_A}\}$ and $B = \{\omega : \omega_1 = b_1, \ldots, \omega_{n_B} = b_{n_B}\}$ be thin cylinders. First, note that if $n \geq n_A$, then

$$P(A)P(B) = \pi(a_1)p(a_1,a_2)\cdots p(a_{n_A-1},a_{n_A})\pi(b_1)p(b_1,b_2)\cdots p(b_{n_B-1},b_{n_B}),$$
$$P(A \cap T^{-n}B) = \pi(a_1)p(a_1,a_2)\cdots p(a_{n_A-1},a_{n_A})p^{(n-n_A+1)}(a_{n_A},b_1)p(b_1,b_2)\cdots p(b_{n_B-1},b_{n_B}),$$

where $p^{(m)}(i,j) = P(X_{1+m} = j | X_1 = i)$. Since the chain is irreducible and aperiodic, $\lim_{n\to\infty} p^{(n-n_A+1)}(a_{n_A}, b_1) = \pi(b_1)$ (for a proof, see [3, p. 125]). Therefore,

$$\lim_{n\to\infty} P(A \cap T^{-n}B) = P(A)P(B),$$

for all thin cylinders A and B. We conclude that T is mixing, and by Theorem 1.9, it is also ergodic.

Example 1.13. Let $X = (X_k)_{k\in\mathbb{N}}$ be a sequence of independent, identically distributed real-valued random variables. In this case, it is easy to show that the shift operator T defined on $(\mathbb{R}^\infty, \mathscr{B}(\mathbb{R}^\infty), P_X)$ is mixing. Let $A = \mathcal{C}(A')$ and $B = \mathcal{C}(B')$ be cylinders with bases A' at $(a_1, a_2, \ldots, a_{n_A})$ and B' at $(b_1, b_2, \ldots, b_{n_B})$. Then, if $n \in \mathbb{N}$ is chosen so large that $b_1 + n > a_{n_A}$, we have

$$\begin{aligned} P_X(A \cap T^{-n}B) &= P[(X_{a_1}, \ldots, X_{a_{n_A}}) \in A', (X_{b_1+n}, \ldots, X_{b_{n_B+n}}) \in B'] \\ &= P[(X_{a_1}, \ldots, X_{a_{n_A}}) \in A']P[(X_{b_1+n}, \ldots, X_{b_{n_B+n}}) \in B'] \\ &= P_X(A)P[(X_{b_1}, \ldots, X_{b_{n_B}}) \in B'] = P_X(A)P_X(B). \end{aligned}$$

Since $\mathscr{B}(\mathbb{R}^\infty)$ is generated by the cylinders, Theorem 1.11 implies that T is mixing.

Example 1.14. The angle doubling transformation we considered before is also mixing. If $A = \{\omega : d_1(\omega) = a_1, \ldots, d_{n_A}(\omega) = a_{n_A}\}$ and B are dyadic intervals, it is easy to check that $P(A \cap T^{-n}B) = P(A)P(B)$ for all $n \geq n_A$. Since $\mathscr{B}(\mathbb{R})$ is generated by the π-system consisting of dyadic intervals and the empty set, Theorem 1.11 again implies that T is mixing.

1.3 Birkhoff's ergodic theorem

Before stating and proving Birkhoff's ergodic theorem, we first prove a preliminary result, the *maximal ergodic theorem*. Its statement and proof is most convenient to express in terms of functional operators.

Let $(\Omega, \mathscr{F}, P)$ be a probability space, and let T be a measure-preserving transformation. Define the operator $U : L^1(\Omega) \to L^1(\Omega)$ by $Uf = f \circ T$, that is,

$$(Uf)(\omega) = f(T\omega),\ f \in L^1(\Omega), \omega \in \Omega.$$

Observe that the operator U is nonnegative in the following sense: if $f \leq g$ (pointwise), then $Uf \leq Ug$.

The fact that T is measure-preserving implies that U preserves expectation:

Lemma 1.15. *For all $f \in L^1(\Omega)$, $E[Uf] = E[f]$.*

Proof. If f is a simple function, then it has representation $f = \sum_{i=1}^n x_i I_{A_i}$, where the x_i are distinct real numbers and the sets A_i are disjoint. But $Uf = \sum_{i=1}^n x_i I_{T^{-1}A_i}$ and thus

$$E[f] = \sum_{i=1}^n x_i P(A_i) = \sum_{i=1}^n x_i P(T^{-1}A_i) = E[Uf].$$

Therefore, the claim is true for simple functions. If f is a nonnegative measurable function, then we can pick an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of simple functions such that $f_n \uparrow f$. But then also $Uf_n \uparrow Uf$, and the monotone convergence theorem implies that

$$E[f] = \lim_{n\to\infty} E[f_n] = \lim_{n\to\infty} E[Uf_n] = E[Uf].$$

Therefore the claim is true for all nonnegative measurable functions. Finally, if $f \in L^1(\Omega)$, then it has decomposition $f = f^+ - f^-$ where f^+ and f^- are nonnegative integrable functions. For every $\omega \in \Omega$, we have

$$(Uf)^+(\omega) = \max\{(Uf)(\omega), 0\} = \max\{f(T\omega), 0\} = f^+(T\omega) = (Uf^+)(\omega),$$

and thus $(Uf)^+ = Uf^+$. Similarly, $(Uf)^- = Uf^-$, and therefore

$$E[f] = E[f^+] - E[f^-] = E[Uf^+] - E[Uf^-] = E[(Uf)^+] - E[(Uf)^-] = E[Uf].$$

□

Define $S_n f = \sum_{i=1}^n U^{i-1} f = \sum_{i=1}^n f \circ T^{i-1}$ for $n \geq 1$, $S_0 f = 0$ (as usual, U^0 is interpreted as identity operator). Put

$$M_n f = \max_{0\leq k\leq n} S_k f,\ M_\infty f = \sup_{n\geq 0} S_n f = \sup_{n\geq 0} M_n f.$$

Then:

Theorem 1.16. *(The maximal ergodic theorem) If $f \in L^1(\Omega)$, then*

$$\int_{\{M_\infty f>0\}} f \,\mathrm{d}P \geq 0.$$

Proof. Put $B_n = \{M_n f > 0\}$, $B_\infty = \{M_\infty f > 0\}$. Now it is clear that $B_n \uparrow B_\infty$ as $n \to \infty$. Suppose that we are able to prove that $\int_{B_n} f\mathrm{d}P \geq 0$ for all $n \geq 1$. Since $f \in L^1(\Omega)$, the dominated convergence theorem implies that

$$\int_{\{M_\infty f>0\}} f \,\mathrm{d}P = \lim_{n\to\infty} \int_{\{M_n f>0\}} f \,\mathrm{d}P \geq 0.$$

Hence, it is enough to show that $\int_{B_n} f\mathrm{d}P \geq 0$ for all $n \geq 1$.

Observe that $(M_n f)I_{B_n} = (\max_{1\leq k\leq n} S_k f)I_{B_n}$. And since U is nonnegative, we have

$$S_k f = f + US_{k-1} f \leq f + UM_n f$$

for all $1 \le k \le n$. These facts imply that $(M_n f) I_{B_n} \le (f + UM_n f) I_{B_n}$. Since $M_n f$ and $UM_n f$ are nonnegative,

$$\begin{aligned}\int_\Omega M_n f \, \mathrm{d}P = \int_{B_n} M_n f \, \mathrm{d}P &\le \int_{B_n} (f + UM_n f) \, \mathrm{d}P \\ &\le \int_{B_n} f \, \mathrm{d}P + \int_\Omega UM_n f \, \mathrm{d}P = \int_{B_n} f \, \mathrm{d}P + \int_\Omega M_n f \, \mathrm{d}P.\end{aligned}$$

The last equality of course follows from Lemma 1.15. Finally, we show that $M_n f$ is integrable:

$$\begin{aligned}|M_n f| = \left| \max_{0 \le k \le n} S_k f \right| &\le \max_{0 \le k \le n} |S_k f| = \max_{0 \le k \le n} \left| \sum_{i=1}^{k} U^{i-1} f \right| \le \max_{0 \le k \le n} \sum_{i=1}^{k} \left| U^{i-1} f \right| \\ &\le \sum_{k=1}^{n} \sum_{i=1}^{k} \left| U^{i-1} f \right| \le \sum_{k=1}^{n} \sum_{i=1}^{k} U^{i-1} |f| \,,\end{aligned}$$

and the right-hand side is integrable by Lemma 1.15. □

Definition 1.17. A measurable function f is *invariant* if $f(T\omega) = f(\omega)$ for all $\omega \in \Omega$. (Observe that, by induction, $f(T^k \omega) = f(\omega)$ for all $k \ge 1$.)

We have now reached the culmination point of this chapter. Here it is:

Theorem 1.18. *(Birkhoff's Ergodic Theorem) Suppose that T is a measure-preserving transformation on a probability space $(\Omega, \mathscr{F}, P)$. If $f \in L^1(\Omega)$, then there exists an invariant and integrable function $\hat{f}$ such that*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(T^{k-1} \omega) = \hat{f}(\omega)$$

with probability 1. Moreover, $E[\hat{f}] = E[f]$, and if T is ergodic, then $\hat{f} = E[f]$ with probability 1.

Remark 1.19. Let A be an invariant set and $f = I_A$. Then clearly

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} f(T^{k-1} \omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \in A^c, \end{cases}$$

and so the limit function $\hat{f}$ can certainly be nonconstant if T is not ergodic.

Proof. Suppose that A is any invariant $\mathscr{F}$-set. Then I_A is an invariant function, and for every $\omega \in \Omega$ we have

$$\begin{aligned}[S_n(fI_A)]\,(\omega) &= \sum_{i=1}^{n}[U^{i-1}(fI_A)](\omega) = \sum_{i=1}^{n}(fI_A)(T^{i-1}\omega) \\ &= \sum_{i=1}^{n} f(T^{i-1}\omega)I_A(T^{i-1}\omega) = \sum_{i=1}^{n} f(T^{i-1}\omega)I_A(\omega) = (S_nf)(\omega)I_A(\omega).\end{aligned}$$

We also have

$$[M_\infty(fI_A)](\omega) = \sup_{n\geq 1}\,[S_n(fI_A)]\,(\omega) = \sup_{n\geq 1}(S_nf)(\omega)I_A(\omega) = (M_\infty f)(\omega)I_A(\omega).$$

Hence, by the maximal ergodic theorem,

$$0 \leq \int_{\{M_\infty(fI_A)>0\}} fI_A\,\mathrm{d}P = \int_{A\cap\{I_AM_\infty f>0\}} f\,\mathrm{d}P = \int_{A\cap\{M_\infty f>0\}} f\,\mathrm{d}P. \tag{1.20}$$

Let $\lambda \in \mathbb{R}$ be a constant. Then

$$\begin{aligned}\{M_\infty(f-\lambda)>0\} &= \bigcup_{n=1}^{\infty}\{S_n(f-\lambda)>0\} = \bigcup_{n=1}^{\infty}\{S_nf - n\lambda > 0\} = \bigcup_{n=1}^{\infty}\left\{\frac{1}{n}S_nf > \lambda\right\} \\ &= \left\{\sup_{n\geq 1}\frac{1}{n}S_nf > \lambda\right\} = \left\{\omega : \sup_{n\geq 1}\frac{1}{n}\sum_{k=1}^{n} f(T^k\omega) > \lambda\right\} := F_\lambda.\end{aligned}$$

By (1.20), $\int_{A\cap F_\lambda}(f-\lambda)\,\mathrm{d}P \geq 0$, or equivalently,

$$\lambda P(A\cap F_\lambda) \leq \int_{A\cap F_\lambda} f\,\mathrm{d}P. \tag{1.21}$$

This holds for all invariant sets A and real numbers λ.

Define $a_n(\omega) = \frac{1}{n}\sum_{k=1}^{n} f(T^{k-1}\omega)$. We want to prove that the a_n converge with probability 1. Let $\alpha < \beta$, and consider the set

$$A_{\alpha,\beta} = \{\omega : \liminf_{n\to\infty} a_n(\omega) < \alpha < \beta < \limsup_{n\to\infty} a_n(\omega)\}.$$

It is clear that

$$\{\omega : a_n(\omega) \text{ does not converge}\} = \bigcup_{\substack{\alpha\in\mathbb{Q},\beta\in\mathbb{Q}\\ \alpha<\beta}} A_{\alpha,\beta}.$$

Hence, it is enough to prove that $P(A_{\alpha,\beta}) = 0$ for all $\alpha < \beta$. Observe that

$$\begin{aligned}\liminf_{n\to\infty} a_n(T\omega) &= \liminf_{n\to\infty} a_{n-1}(T\omega) = \liminf_{n\to\infty} \frac{1}{n-1}\sum_{k=1}^{n-1} f(T^{k-1}T\omega) \\ &= \liminf_{n\to\infty} \frac{1}{n-1}\sum_{k=1}^{n} f(T^{k-1}\omega) - \frac{f(\omega)}{n-1} = \liminf_{n\to\infty} \frac{n}{(n-1)n}\sum_{k=1}^{n} f(T^{k-1}\omega) \\ &= \liminf_{n\to\infty} a_n(\omega).\end{aligned}$$

Similarly, $\limsup_{n\to\infty} a_n(T\omega) = \limsup_{n\to\infty} a_n(\omega)$, which implies that $T^{-1}A_{\alpha,\beta} = A_{\alpha,\beta}$. Since $A_{\alpha,\beta} = A_{\alpha,\beta} \cap F_\beta$, (1.21) implies that

$$\beta P(A_{\alpha,\beta}) = \beta P(A_{\alpha,\beta} \cap F_\beta) \le \int_{A_{\alpha,\beta}\cap F_\beta} f \,\mathrm{d}P = \int_{A_{\alpha,\beta}} f \,\mathrm{d}P. \tag{1.22}$$

And since $-\beta < -\alpha$, applying the same reasoning to $-f$ gives

$$-\alpha P(A^-_{-\beta,-\alpha}) \le \int_{A^-_{-\beta,-\alpha}} -f \,\mathrm{d}P, \tag{1.23}$$

where $A^-_{-\beta,-\alpha}$ is the analogue of $A_{\alpha,\beta}$ for $-f$. But

$$\begin{aligned}A^-_{-\beta,-\alpha} &= \{\omega : \liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} -f(T^{k-1}\omega) < -\beta < -\alpha < \limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} -f(T^{k-1}\omega)\} \\ &= \{\omega : -\limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} -f(T^{k-1}\omega) < \alpha < \beta < -\liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} -f(T^{k-1}\omega)\} \\ &= \{\omega : \liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(T^{k-1}\omega) < \alpha < \beta < \limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} f(T^{k-1}\omega)\} \\ &= A_{\alpha,\beta}.\end{aligned}$$

Therefore, (1.23) and (1.22) imply that $\alpha P(A_{\alpha,\beta}) \ge \int_{A_{\alpha,\beta}} f \,\mathrm{d}P \ge \beta P(A_{\alpha,\beta})$. But since $\alpha < \beta$, this is possible if and only if $P(A_{\alpha,\beta}) = 0$.

We have shown that $\frac{1}{n}\sum_{k=1}^{n} f(T^{k-1}\omega) = a_n(\omega)$ converges on a set with probability 1. Define $g(\omega) = \lim_{n\to\infty} a_n(\omega)$ on the set where a_n converges, and let $g(\omega) = 0$ elsewhere. Observe that $g(\omega)$ may assume the values ∞ and $-\infty$ at certain values of ω. We will show that $E[g] = E[f]$, which implies that $|g(\omega)| < \infty$ almost surely.

By Lemma 1.15,

$$E[a_n] = \frac{1}{n}\sum_{k=1}^{n} E[f \circ T^{k-1}] = \frac{1}{n}\sum_{k=1}^{n} E[f] = E[f]$$

for all $n \geq 1$. If we can show that the functions a_n are uniformly integrable, then it follows from Theorem 0.27 that $E[f] = \lim_{n\to\infty} E[a_n] = E[g]$ and g is integrable (and thus finite almost surely).

Let λ be an arbitrary nonnegative real number. By (1.21), $\lambda P(F_\lambda) = \lambda P(\Omega \cap F_\lambda) \leq \int_{\Omega\cap F_\lambda} f \,\mathrm{d}P = \int_{F_\lambda} f \,\mathrm{d}P \leq E[|f|] < \infty$. If $G_\lambda = \{\sup_{n\geq 1} |a_n| > \lambda\}$, then

$$G_\lambda = \{\sup_{n\geq 1} a_n > \lambda\} \cup \{\sup_{n\geq 1} a_n < -\lambda\} = F_\lambda \cup \{\sup_{n\geq 1} -a_n > \lambda\}.$$

But the set $\{\sup_{n\geq 1} -a_n > \lambda\}$ is the analogue of F_λ for $-f$. Again by (1.21),

$$\lambda P(\{\sup_{n\geq 1} -a_n > \lambda\}) \leq \int_{\{\sup_{n\geq 1} -a_n > \lambda\}} -f \,\mathrm{d}P \leq E[|f|] < \infty,$$

and we conclude that $\lambda P(G_\lambda) \leq 2E[|f|]$. If α and λ are positive, then

$$\begin{aligned}
\int_{\{|a_n|>\lambda\}} |a_n| \,\mathrm{d}P &\leq \int_{G_\lambda} |a_n| \,\mathrm{d}P \leq \frac{1}{n}\sum_{k=1}^{n} \int_{G_\lambda} |f\circ T^{k-1}| \,\mathrm{d}P \\
&= \frac{1}{n}\sum_{k=1}^{n} \int_{(G_\lambda\cap\{|f\circ T^{k-1}|>\alpha\})\cup(G_\lambda\cap\{|f\circ T^{k-1}|\leq\alpha\})} |f\circ T^{k-1}| \,\mathrm{d}P \\
&\leq \frac{1}{n}\sum_{k=1}^{n} \left(\int_{\{|f\circ T^{k-1}|>\alpha\}} |f\circ T^{k-1}| \,\mathrm{d}P + \alpha P(G_\lambda) \right).
\end{aligned}$$

But $\int_{\{|f\circ T^{k-1}|>\alpha\}} |f\circ T^{k-1}| \,\mathrm{d}P = \int_\Omega U^{k-1}\left[(I_{(\alpha,\infty]}\circ|f|)|f|\right] \,\mathrm{d}P = \int_\Omega (I_{(\alpha,\infty]}\circ|f|)|f| \,\mathrm{d}P = \int_{\{|f|>\alpha\}} |f| \,\mathrm{d}P$ by Lemma 1.15. Hence,

$$\int_{\{|a_n|>\lambda\}} |a_n| \,\mathrm{d}P \leq \int_{\{|f|>\alpha\}} |f| \,\mathrm{d}P + \alpha P(G_\lambda) \leq \int_{\{|f|>\alpha\}} |f| \,\mathrm{d}P + 2\frac{\alpha}{\lambda} E[|f|].$$

Put $\alpha = \sqrt{\lambda}$. Then, if $\lambda \to \infty$, the final expression goes to zero since f is integrable. We may conclude that the a_n are uniformly integrable.

Now since $E[g] = E[f] < \infty$, $\lim_{n\to\infty} a_n$ exists and is finite on a set with probability 1. Define $\hat{f}(\omega) = \lim_{n\to\infty} a_n(\omega)$ on this set, and let $\hat{f}(\omega) = 0$ elsewhere. Then $\hat{f} = g$ almost surely and so $E[\hat{f}] = E[g] = E[f]$. Since $\liminf_{n\to\infty} a_n(\omega) = \liminf_{n\to\infty} a_n(T\omega)$ and $\limsup_{n\to\infty} a_n(\omega) = \limsup_{n\to\infty} a_n(T\omega)$, we have $\hat{f}(\omega) = \hat{f}(T\omega)$: $\hat{f}$ is invariant as proposed.

Finally, suppose that T is ergodic. Observe that the set $\{\omega : \hat{f}(\omega) \leq x\}$ is invariant, which implies that its probability is either 0 or 1. Let x_0 be the infimum of the x for which

it is 1. By the well-known properties of cumulative distribution functions, $P(\hat{f} \le x_0) = 1$ and

$$P(\hat{f} = x_0) = P(\hat{f} \le x_0) - P(\hat{f} < x_0) = 1 - \lim_{x \to x_0^-} P(\hat{f} \le x) = 1 - 0 = 1.$$

Therefore, $\hat{f}$ is constant x_0 almost surely and thus $x_0 = E[\hat{f}] = E[f]$. We conclude that in the ergodic case, $\hat{f} = E[f]$ with probability 1. This completes the proof. □

It is easy to check that the collection of invariant $\mathscr{F}$-sets forms a σ-algebra. Let this σ-algebra be denoted by $\mathscr{I}$. The function $\hat{f}$ will now be identified as the conditional expectation of f given $\mathscr{I}$.

If G is any invariant set, then

$$\begin{aligned}\int_G \frac{1}{n}\sum_{k=1}^{n} f(T^{k-1}\omega)\,P(\mathrm{d}\omega) &= \frac{1}{n}\sum_{k=1}^{n}\int_\Omega f(T^{k-1}\omega) I_G(\omega)\,P(\mathrm{d}\omega)\\ &= \frac{1}{n}\sum_{k=1}^{n}\int_\Omega f(T^{k-1}\omega) I_G(T^{k-1}\omega)\,P(\mathrm{d}\omega)\\ &= \frac{1}{n}\sum_{k=1}^{n} E[U^{k-1}(fI_G)] = \frac{1}{n}\sum_{k=1}^{n} E[fI_G] = E[fI_G].\end{aligned}$$

But since the averages $\frac{1}{n}\sum_{k=1}^{n} f(T^{k-1}\omega)$ converge to $\hat{f}(\omega)$ almost surely and they are uniformly integrable, we have

$$\int_G \hat{f}\,\mathrm{d}P = \lim_{n\to\infty}\int_G \frac{1}{n}\sum_{k=1}^{n} f(T^{k-1}\omega)\,P(\mathrm{d}\omega) = \int_G f\,\mathrm{d}P.$$

Therefore, $E[f|\mathscr{I}] = \hat{f}$ by the definition of conditional expectation.

Example 1.24. Let $\Omega = \{a, b, c, d, e\}$, and let $T = (abc)(de)$, a product of two cycles. Let equal probabilities be given to a, b, c and d, e so that T is measure-preserving. If $A = \{a, d\}$ and $f = I_A$, then the limit function $\hat{f}$ is $\frac{1}{3}$ on $\{a, b, c\}$ and $\frac{1}{2}$ on $\{d, e\}$.

Example 1.25. Let us now use the ergodic theorem to prove an interesting fact about the unit interval: for almost every number, the proportion of ones in the binary expansion up to the nth digit tends to $\frac{1}{2}$ as $n \to \infty$, that is,

$$P\left(\left\{\omega : \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} d_k(\omega) = \frac{1}{2}\right\}\right) = 1,$$

where $d_k(\omega)$ is the kth digit of the nonterminating binary expansion of ω.

Let T be the angle doubling transformation on $((0,1], \mathscr{B}((0,1]), m)$, where m is the Lebesgue measure. As we have already proved, T shifts the binary digits of ω to the left: $d_k(T\omega) = d_{k+1}(\omega)$ for all $\omega \in (0,1]$ and $k \in \mathbb{N}$. Therefore,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} d_k(\omega) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} d_1(T^{k-1}\omega).$$

Since T is ergodic and $E[d_1] = \frac{1}{2}$, the ergodic theorem then implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} d_k(\omega) = E[d_1] = \frac{1}{2}$$

with probability 1.

1.4 Ergodic stochastic processes

Ergodicity of a stochastic process is defined in terms of its distribution:

Definition 1.26. Let $T = \mathbb{Z}$ or $T = \mathbb{N}$. We say that a stochastic process $X = (X_k)_{k\in T}$ with state space $(S, \mathscr{S})$ is ergodic, if the shift transformation on $(S^T, \mathscr{S}^T, P_X)$ is ergodic.

Suppose that the shift transformation T' on S^T (the usual labeling T is now reserved for the parameter set) is indeed ergodic, and further suppose that X is stationary so that T' is also measure-preserving by Theorem 1.5. If $f : S^T \to \mathbb{R}$ is measurable $\mathscr{S}^T/\mathscr{B}(\mathbb{R})$ and integrable, then the ergodic theorem implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(T'^{k-1}x) = \int_{S^T} f(x)\, P_X(\mathrm{d}x) \tag{1.27}$$

for all x on a set of P_X-measure 1. By Lemma 0.13 we have

$$\int_{S^T} f(x)\, P_X(\mathrm{d}x) = \int_{\Omega} f(X(\omega))\, P(\mathrm{d}\omega).$$

Since (1.27) holds with P_X-probability 1, we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(T'^{k-1}X(\omega)) = \int_{\Omega} f(X(\omega))\, P(\mathrm{d}\omega) = E[f(X)] \tag{1.28}$$

with P-probability 1. To see this, simply note that if $A \subset S^T$ is the set on which (1.27) holds, then $1 = P_X(A) = P(\omega : X(\omega) \in A)$. Note also that if $T = \mathbb{N}$, then (1.28) becomes

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k(\omega), X_{k+1}(\omega), \ldots) = E[f(X_1, X_2, \ldots)]. \tag{1.29}$$

Example 1.30. The Strong Law of Large Numbers is an immediate consequence of the ergodic theorem. Let $(X_k)_{k\in\mathbb{N}}$ be an independent sequence of identically distributed real-valued random variables with finite expectation m. Then the shift transformation T on $\mathbb{R}^\infty$ is measure-preserving and ergodic as we have previously concluded. Let $f : \mathbb{R}^\infty \to \mathbb{R}$ be the first coordinate function, that is, $f(x_1, x_2, \ldots) = x_1$. Then f is certainly Borel measurable, and since

$$\int_{\mathbb{R}^\infty} f(x)\, P_X(\mathrm{d}x) = \int_\Omega f(X)\, \mathrm{d}P = \int_\Omega X_1\, \mathrm{d}P = m < \infty,$$

it is integrable. By (1.29) we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} X_k(\omega) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k(\omega), X_{k+1}(\omega), \ldots) = E[f(X_1, X_2, \ldots)] = E(X_1) = m$$

with probability 1.

Example 1.31. Suppose that $(X_k)_{k\in\mathbb{N}}$ is an irreducible, aperiodic Markov chain with finite state space S and stationary distribution π which coincides with the initial distribution of the process. Again, the shift transformation T on $(S^\infty, (\mathcal{P}(S))^\infty, P_X)$ is measure-preserving and ergodic as we have previously seen. If $j \in S$ and $f(x) = I_{\{j\}}(x_1)$, $x \in S^\infty$, then f is clearly Borel measurable and integrable. By (1.29) we have

$$\begin{aligned}
\lim_{n\to\infty} \frac{\#\{1 \le k \le n : X_k(\omega) = j\}}{n} &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k(\omega), X_{k+1}(\omega), \ldots) \\
&= E[f(X_1, X_2, \ldots)] = E[I_{\{j\}}(X_1)] = P(X_1 = j) \\
&= \pi(j)
\end{aligned}$$

with probability 1.

Chapter 2

Shannon-McMillan-Breiman Theorem

In this chapter we will apply the ergodic theorem to prove a famous result in information theory, the Shannon-McMillan-Breiman theorem.

All random variables in this chapter will be discrete. This means that if $X : \Omega \to S$ is a random variable, then the *probability mass function* $p_X : S \to \mathbb{R}$ defined by $p_X(x) = P(X = x)$ satisfies $\sum_{x \in S} p_X(x) = 1$. This of course implies that the set on which p_X is positive is at most countable. Usually $p_X(x)$ will be simply denoted by $p(x)$ if it is clear from context that p is the probability mass function of X. Similarly, if $X_1, X_2, \ldots, X_n$ are discrete random variables, then the value of the joint probability mass function $p_{X_1,X_2,\ldots,X_n}$ at point $(x_1, x_2, \ldots, x_n)$ is denoted by $p(x_1, x_2, \ldots, x_n)$.

We will apply similar notation for conditional probabilities as well. If X and Y are random variables, then $P(X = x|Y = y)$ is denoted by $p_{X|Y}(x|y)$ or $p(x|y)$ if the meaning is clear from context. Similarly, if $X_1, X_2, \ldots, X_n$ are discrete random variables, then $p(x_n|x_{n-1}, x_{n-2}, \ldots, x_1)$ means $P(X_n = x_n \mid X_{n-1} = x_{n-1}, \ldots, X_1 = x_1)$, and so on.

We will also often consider random variables such $p(X)$, which of course means the random variable that maps ω to $p_X(X(\omega))$. Conditional probabilities such as $P(X_n = x_n \mid X_{n-1}, \ldots, X_1)$ may also be written as $p(x_n \mid X_{n-1}, \ldots, X_1)$.

2.1 Basic concepts of Information Theory

We begin this chapter with a brief introduction to information theory. The most fundamental quantity in information theory is called *entropy*:

Definition 2.1. Entropy of a discrete random variable X with probability mass function $p(x)$ is defined as

$$H(X) = -\sum_{x \in S} p(x) \log_2(p(x)) = -E[\log_2 p(X)].$$

In information theory, it is customary to use base 2 logarithms. From here on, log always means base 2 logarithm unless stated otherwise. We also define $0 \log 0 = 0$, which is justified by the fact that $\lim_{x \to 0^+} x \log x = 0$. Thus we don't have to assume that $p(x) > 0$ for all $x \in S$.

Observe that entropy depends only on the probabilities $p(x), x \in S$, but not on the actual values that X assumes. It is also worth noting that entropy always exists, since the summands are always negative. But entropy may well be infinite.

Definition 2.2. Suppose that $X_1, X_2, \ldots, X_n$ are discrete random variables such that X_k takes values in S_k, $1 \leq k \leq n$. Then the *joint entropy* of $X_1, X_2, \ldots, X_n$ is defined as

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= - \sum_{x_1 \in S_1, \ldots, x_n \in S_n} p(x_1, x_2, \ldots, x_n) \log p(x_1, x_2, \ldots, x_n) \\ &= -E[\log p(X_1, X_2, \ldots, X_n)]. \end{aligned}$$

Observe that since the the random variables $X_1, X_2, \ldots, X_n$ can be treated as a single random vector $(X_1, X_2, \ldots, X_n)$ taking values in $S_1 \times \cdots \times S_n$, nothing new is actually involved here. The joint entropy of $X_1, X_2, \ldots, X_n$ clearly equals the entropy of $(X_1, X_2, \ldots, X_n)$.

Definition 2.3. Suppose that $X : \Omega \to S_X$ and $Y : \Omega \to S_Y$ are discrete random variables. Then the *conditional entropy of Y given $X = x$* is defined as

$$H(Y \mid X = x) = - \sum_{y \in S_Y} p(y|x) \log p(y|x),$$

and the *conditional entropy of Y given X*, denoted by $H(Y|X)$, is defined as the weighted average of the $H(Y \mid X = x)$, that is,

$$H(Y|X) = \sum_{x \in S_X} H(Y \mid X = x) p(x) = - \sum_{x \in S_X, y \in S_Y} p(x, y) \log p(y|x) = -E[\log p(Y|X)].$$

Claude E. Shannon, who laid the foundations of information theory, called the number $H(X)$ entropy since he recognized some analogies between it and the concept of entropy in statistical mechanics. Entropy can be seen as a measure of uncertainty associated with a random variable[5, p. 3]. This is illustrated by the following example.

Example 2.4. Let X be a Bernoulli(p) distributed random variable such that $P(X = 1) = p$ and $P(X = 0) = 1 - p$. In this case,

$$H(X) = -p(\log p) - (1 - p) \log(1 - p).$$

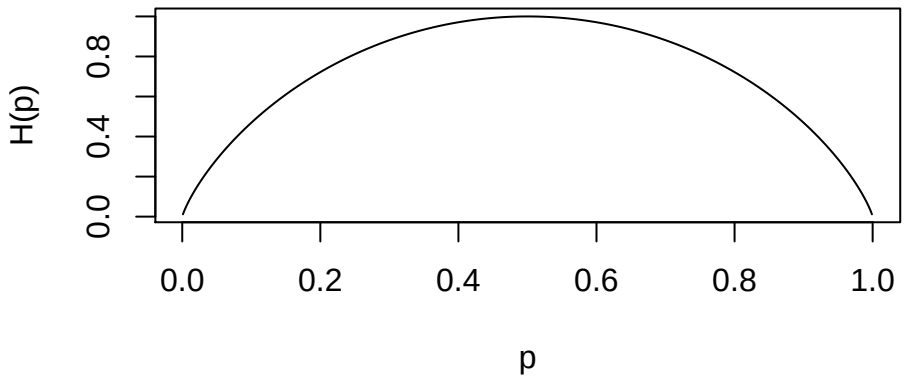

Figure 2.1: The graph of binary entropy function H(p)

This entropy as a function of p, denoted by $H(p)$, is called *the binary entropy function*. As the figure shows, the binary entropy function attains its maximum value 1 at $p = 0.5$, and is zero at $p = 0$ and $p = 1$. This makes a lot of sense if entropy is interpreted as a measure of uncertainty. If p is zero or one, there is no uncertainty and thus entropy is zero as one would expect. On the other hand, if we toss a fair coin (so that $p = 0.5$), more uncertainty concerning the outcome is involved than in the case of a weighted coin.

Example 2.5. Let X denote the number of heads before the first tail in a fair coin tossing. Then $P(X = k) = \frac{1}{2^{k+1}}$ and

$$H(X) = -\sum_{k=0}^{\infty} \frac{1}{2^{k+1}} \log \frac{1}{2^{k+1}} = -\sum_{k=0}^{\infty} \frac{1}{2^{k+1}}(-k-1) = \frac{1}{2}\sum_{k=0}^{\infty}(k+1)\left(\frac{1}{2}\right)^k$$
$$= \frac{1}{2}\sum_{k=1}^{\infty} k\left(\frac{1}{2}\right)^{k-1} = \frac{1}{2}\left(\frac{1}{\left(1-\frac{1}{2}\right)^2}\right) = 2.$$

Kullback-Leibler distance and mutual information

Before we prove the basic properties of entropy, we will briefly discuss the concepts of Kullback-Leibler distance and mutual information. Their nonnegativity will turn out extremely useful in the proofs.

Definition 2.6. Suppose that p and q are probability mass functions on a set S. Then the *Kullback-Leibler distance* between p and q, denoted by $D(p||q)$, is defined as

$$D(p||q) = \sum_{x \in S} p(x) \log \frac{p(x)}{q(x)}.$$

Remark 2.7. We adapt the conventions here: $0 \log \frac{0}{0} = 0, 0 \log \frac{0}{c} = 0$ and $c \log \frac{c}{0} = \infty$. These are all justified by considering appropriate limits.

Even though $D(p||q)$ is called distance between p and q, it is not a metric since it does not fulfill the triangle equality and it is not symmetric. However, the next theorem shows that $D(p||q) = 0$ if and only if $p = q$ and $D(p||q)$ is always nonnegative. Later, many proofs will be based on this important fact.

Theorem 2.8. *Let p and q be probability mass functions on a set S. Then*

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all $x \in S$.

Proof. If for some $x \in S$ we have $p(x) > 0$ and $q(x) = 0$, then $D(p||q) = \infty > 0$ by the convention that $c \log \frac{c}{0} = \infty$. Therefore, we may assume that such x does not exist.

Let $S' = \{x \in S : p(x) > 0\}$ be the support of p. Suppose that X is any random variable with probability mass function p. Then

$$\begin{aligned} -D(p||q) &= -\sum_{x \in S} p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in S'} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x \in S'} p(x) \log \frac{q(x)}{p(x)} = E\left[\log \frac{q(X)}{p(X)}\right] = -E\left[-\log \frac{q(X)}{p(X)}\right]. \end{aligned}$$

The mapping $x \mapsto -\log x$ is strictly convex on $(0, \infty)$ and $P\left(\frac{q(X)}{p(X)} > 0\right) = 1$. Thus, by applying Jensen's inequality, a well-known result in probability theory, we obtain

$$-\log\left(E\left[\frac{q(X)}{p(X)}\right]\right) \leq E\left[-\log \frac{q(X)}{p(X)}\right],$$

or equivalently, $-E\left[-\log \frac{q(X)}{p(X)}\right] \leq \log\left(E\left[\frac{q(X)}{p(X)}\right]\right)$. Therefore,

$$\begin{aligned} -D(p||q) &= -E\left[-\log \frac{q(X)}{p(X)}\right] \leq \log\left(E\left[\frac{q(X)}{p(X)}\right]\right) = \log \sum_{x \in S'} p(x) \frac{q(x)}{p(x)} \\ &= \log \sum_{x \in S'} q(x) \leq \log \sum_{x \in S} q(x) = \log 1 = 0. \end{aligned}$$

This proves that Kullback-Leibler distance is nonnegative.

What remains to be proven is that we have equality if and only if $p(x) = q(x)$ for all $x \in S$. Suppose first that we have $D(p||q) = 0$. Then the two inequalities above must actually be equalites. By the strict convexity of the mapping $x \mapsto -\log x$, this is possible in Jensen's inequality only if the random variable $\frac{q(X)}{p(X)}$ is constant almost surely. This is equivalent to having $\frac{q(x)}{p(x)} = c$ for some real number c and for all $x \in S'$. Summing over all $x \in S'$ we obtain

$$\sum_{x \in S'} q(x) = c \sum_{x \in S'} p(x) = c.$$

But we must also have equality in $\log \sum_{x \in S'} q(x) \leq \log \sum_{x \in S} q(x)$, which implies $c = \sum_{x \in S'} q(x) = \sum_{x \in S} q(x) = 1$. Thus $c = 1$ and we have $p(x) = q(x)$ for all $x \in S'$. Now having $\sum_{x \in S'} q(x) = \sum_{x \in S} q(x)$ further implies that $p(x) = q(x)$ for all all $x \in S$.

Conversely, if $p(x) = q(x)$ for all $x \in S$, then we have

$$\log \sum_{x \in S} q(x) = \log \sum_{x \in S} p(x) = \log \sum_{x \in S'} p(x) = \log \sum_{x \in S'} q(x).$$

Also the random variable $\frac{q(X)}{p(X)}$ is constant almost surely, which implies that we also have equality in Jensen's inequality. Thus $D(p||q) = 0$. □

Definition 2.9. The *mutual information* of discrete random variables X and Y, denoted by $I(X;Y)$, is the Kullback-Leibler distance between the joint probability mass functions $p_{X,Y}(x,y)$ and $p_X(x)p_Y(y)$ defined on $S_X \times S_Y$.

Mutual information measures how much information one random variable contains about another random variable. This is clarified by the fact that if X and Y are independent, then their product distribution equals their joint distribution and we have $I(X;Y) = 0$ by the previous theorem. The theorem also states that we always have $I(X;Y) \geq 0$ with equality if and only if X and Y are independent.

Let us derive a handy formula for the mutual information $I(X;Y)$:

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\ &= -\sum_{x,y} p(x,y) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y)) \\ &= -\sum_{x} p(x) \log p(x) - \left(-\sum_{x,y} p(x,y) \log p(x|y) \right) \\ &= H(X) - H(X|Y). \end{aligned}$$

And since $I(X;Y) \geq 0$, we have the following inequality:

$$H(X) \geq H(X|Y). \tag{2.10}$$

Properties of entropy

We will now prove the basic properties of entropy. We start with the following theorem which illustrates the relationshop between joint and conditional entropy:

Theorem 2.11. *If $X : \Omega \to S_X$ and $Y : \Omega \to S_Y$ are discrete random variables, then*

$$H(X,Y) = H(X) + H(Y \mid X).$$

Proof.

$$\begin{aligned} H(X,Y) &= -\sum_{x\in S_X}\sum_{y\in S_Y} p(x,y)\log p(x,y) = -\sum_{x\in S_X}\sum_{y\in S_Y} p(x,y)\log\left[p(x)p(y|x)\right] \\ &= -\sum_{x\in S_X}\sum_{y\in S_Y} p(x,y)\log p(x) - \sum_{x\in S_X}\sum_{y\in S_Y} p(x,y)\log p(y|x) \\ &= -\sum_{x\in S_X} p(x)\log p(x) - \sum_{x\in S_X}\sum_{y\in S_Y} p(x,y)\log p(y|x) \\ &= H(X) + H(Y \mid X). \end{aligned}$$

□

Corollary 2.12. *Suppose that $X_1, X_2, \ldots, X_{n-1}$ and X_n are discrete random variables and $n \geq 2$. Then*

$$H(X_1, X_2, \ldots, X_n) = \sum_{k=1}^{n} H(X_k \mid X_{k-1}, X_{k-2}, \ldots, X_1).$$

Proof. We prove the claim by induction. By the theorem, the claim is true if $n = 2$. Suppose then that it holds for $n-1$ random variables. In this case,

$$\begin{aligned} H(X_1, X_2, \ldots, X_n) &= H\left[(X_1, \ldots, X_{n-1}), X_n\right] \\ &= H(X_1, \ldots, X_{n-1}) + H(X_n \mid X_{n-1}, \ldots, X_1) \\ &= \sum_{k=1}^{n-1} H(X_k \mid X_{k-1}, X_{k-2}, \ldots, X_1) + H(X_n \mid X_{n-1}, \ldots, X_1) \\ &= \sum_{k=1}^{n} H(X_k \mid X_{k-1}, X_{k-2}, \ldots, X_1). \end{aligned}$$

By induction, the claim holds for all $n \geq 2$. □

Theorem 2.11 has an analogue for conditional entropy:

Theorem 2.13. *If $X : \Omega \to S_X, Y : \Omega \to S_Y$ and $Z : \Omega \to S_Z$ are discrete random variables, then*

$$H(X, Y \mid Z) = H(X \mid Z) + H(Y \mid X, Z).$$

Proof. Since $p_{X,Y|Z}(x, y|z) = p_{X|Z}(x|z)p_{Y|X,Z}(y|x, z)$ whenever $p_{X,Z}(x, z) > 0$, we have

$$\begin{aligned} H(X, Y \mid Z) &= -\sum_{x,y,z} p(x, y, z) \log p(x, y|z) = -\sum_{x,y,z} p(x, y, z) \log \left[p(x|z)p(y|x, z)\right] \\ &= -\sum_{x,y,z} p(x, y, z) \log p(x|z) - \sum_{x,y,z} p(x, y, z) \log p(y|x, z) \\ &= -\sum_{x,z} p(x, z) \log p(x|z) - \sum_{x,y,z} p(x, y, z) \log p(y|x, z) \\ &= H(X \mid Z) + H(Y \mid X, Z). \end{aligned}$$

□

Recall that always $H(X) \geq H(X|Y)$. As an important consequence of this fact we have the following theorem:

Theorem 2.14. *Suppose that $X_1, X_2, \ldots, X_{n-1}$ and X_n are discrete random variables. Then*

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{k=1}^{n} H(X_i)$$

with equality if and only if the random variables are independent.

Proof. By Theorem 2.12,

$$H(X_1, X_2, \ldots, X_n) = \sum_{k=1}^{n} H(X_k|X_{k-1}, X_{k-2}, \ldots, X_1) \leq \sum_{k=1}^{n} H(X_k).$$

We have equality here if and only if $H(X_k|X_{k-1}, X_{k-2}, \ldots, X_1)$ equals $H(X_k)$ for each k. But this happens if and only if X_k is independent of $(X_{k-1}, \ldots, X_1)$ for each k. □

Corollary 2.15. *If the X_i are also identically distributed, then $H(X_1, X_2, \ldots, X_n) = nH(X_1)$.*

An analogous result also holds for conditional entropy:

Theorem 2.16. *Suppose that $X_1, X_2, \ldots, X_n$ and Z are discrete random variables. Then*

$$H(X_1, X_2, \ldots, X_n \mid Z) \leq \sum_{i=1}^{n} H(X_i \mid Z)$$

with equality if and only if the random variables X_k are conditionally independent given Z, that is,

$$p_{X_i, X_j|Z}(i, j|z) = p_{X_i}(i|z) p_{X_j}(j|z)$$

for all i, j and z.

Proof. Let $i \in S_Z$. Then by Theorem 2.14,

$$H(X_1, X_2, \ldots, X_n | Z = i) \leq \sum_{k=1}^{n} H(X_k | Z = i)$$

with equality if and only if the X_k are independent given $Z = i$. The claim follows by multiplying this inequality by $p_Z(i)$ and summing over all $i \in S_Z$. □

The analogue of inequality (2.10) for conditional entropy is given by the following theorem:

Theorem 2.17. *If X, Y and Z are discrete random variables, then*

$$H(Z \mid X, Y) \leq H(Z \mid X)$$

with equality if and only if Y and Z are conditionally independent given X.

Proof. By Theorems 2.13 and 2.16, we have

$$\begin{aligned} H(Z \mid X, Y) &= H(Y, Z \mid X) - H(Y \mid X) \leq H(Y, Z \mid X) - H(Y, Z \mid X) + H(Z \mid X) \\ &= H(Z \mid X) \end{aligned}$$

with equality if and only if Y and Z are conditionally independent given X. □

Observe that, informally, this theorem and inequality (2.10) state that conditioning always reduces entropy.

2.2 Entropy and stochastic processes

Consider a sequence $(X_k)_{k\in\mathbb{N}}$ of independent and identically distributed random variables. Unless $H(X_1) = 0$, we have $\lim_{n\to\infty} H(X_1, H_2, \ldots, H_n) = \lim_{n\to\infty} nH(X_1) = \infty$. However, $\lim_{n\to\infty} \frac{1}{n} H(X_1, H_2, \ldots, H_n) = H(X_1)$. This justifies the following definition:

Definition 2.18. The *entropy rate* of a stochastic process $X = (X_k)_{k\in T}$ with parameter set $T = \mathbb{N}$ or $T = \mathbb{Z}$ is defined by

$$H(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

when the limits exists.

Example 2.19. Consider a sequence $X = (X_k)_{k\in\mathbb{N}}$ of independent Bernoulli trials such that $p_k = P(X_k = 1)$ is not constant. In this case, $H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i)$. Let

$$p_i = \begin{cases} 0.5 & \text{if } 2k < \log\log i \leq 2k+1, \\ 0 & \text{if } 2k+1 < \log\log i \leq 2k+2 \end{cases}$$

for $k = 0, 1, 2, \ldots$. Now $H(X_i) = H(0.5) = 1$ for arbitrarily long segments, and these are followed by exponentially longer segments where $H(X_i) = H(0) = 0$. Then again, we have an exponentially longer segment with $H(X_i) = 1$, and so on. Hence, the average $\frac{1}{n}\sum_{i=1}^{n} H(X_i)$ oscillates between zero and one. Entropy rate is thus not defined for this process.

The next theorem shows that for stationary processes with $H(X_1) < \infty$, the entropy rate always exists. We also have, in the stationary case, an alternative and often easier formula for calculating the entropy rate.

Theorem 2.20. *If $X = (X_k)_{k\in T}$ is stationary and $H(X_1) < \infty$, then*

$$\lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$$

exists, is finite, and equals $H(X)$.

Proof. By Theorem 2.17 and stationarity,

$$H(X_{n+1} \mid X_n, \ldots, X_1) \leq H(X_{n+1} \mid X_n, \ldots, X_2) = H(X_n \mid X_{n-1}, \ldots, X_1).$$

Since $H(X_{n+1} \mid X_n, \ldots, X_1), n \in \mathbb{N}$, is a decreasing sequence of nonnegative numbers, it converges to a limit. And since $H(X_n \mid X_{n-1}, \ldots, X_1) \leq H(X_1) < \infty$, the limit is finite. Denote this limit by H'. We will show that $H' = H(X)$.

Note that if a_n is a sequence of real numbers such that $\lim_{n\to\infty} a_n = a$, then also $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} a_i = a$. By Corollary 2.12,

$$\frac{H(X_1, X_2, \ldots, X_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1).$$

Therefore,

$$\begin{aligned} H' &= \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(X_i \mid X_{i-1}, \ldots, X_1) \\ &= \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} = H(X). \end{aligned}$$

□

Example 2.21. Consider an aperiodic, irreducible Markov chain $X = (X_k)_{k\in\mathbb{N}}$ with stationary distribution π, transition probabilities $p(i,j)$ and finite state space S. Again, we suppose that the initial distribution is π and thus the process is stationary. Now

$$\begin{aligned} H(X) &= \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) = \lim_{n\to\infty} H(X_n \mid X_{n-1}) = H(X_2|X_1) \\ &= \sum_{i\in S} \pi(i) H(X_2|X_1 = i) = -\sum_{i\in S}\sum_{j\in S} \pi(i)p(i,j)\log p(i,j). \end{aligned}$$

The significance of the entropy rate of a stochastic process should become clear in the next section.

2.3 Asymptotic Equipartition Property

Suppose that a weighted coin with $P(\text{"head"}) = 0.8$ is tossed 1000 times, and suppose further that this experiment is repeated, say, 1000 times. Thus, we obtain 1000 sequences consisting of 1000 heads or tails. It is intuitively clear that most of these sequences contain around 800 heads. The probability of observing one such a sequence is close to $0.8^{800}(1-0.8)^{1000-800}$ which can be written as

$$2^{800\log(0.8)+(1000-800)\log(1-0.8))} = 2^{-1000H(0.8)}.$$

Recall that $H(0.8)$ is the entropy associated with a coin tossing with weight 0.8. It is also the entropy rate associated with the stochastic process defined by this random experiment.

An analogous result is true more generally. If a stochastic process X satisfies certain assumptions which will be discussed shortly, the probability of observing a sequence

$(x_1, x_2, \ldots, x_n)$ is arbitrarily close to $2^{-nH(X)}$ for most of the sequences as n grows. This allows us to partition the space of all sequences of length n into two groups: the typical sequences $(x_1, x_2, \ldots, x_n)$ with $p(x_1, x_2, \ldots, x_n)$ close to $2^{-nH(X)}$, and atypical sequences.

Processes for which this is possible have the *Asymptotic Equipartition Property*. We begin this section by proving the AEP for independent, identically distributed sequences.

Theorem 2.22. *If $X_1, X_2, \ldots$ are independent, identically distributed random variables such that $H(X_1) < \infty$, then*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X_1) \text{ in probability as } n \to \infty.$$

Proof. The theorem is a direct consequence of the weak law of large numbers. Since the X_i are independent and identically distributed, so are the random variables $-\log p(X_i)$. For each i, we have $E[-\log p(X_i)] = H(X_1)$ and thus

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n}\log\left[p(X_1)p(X_2)\cdots p(X_n)\right] = \frac{1}{n}\sum_{i=1}^{n} -\log p(X_i)$$

converges in probability to $E[-\log p(X_i)] = H(X_1)$ by the weak law of large numbers. □

Definition 2.23. Let $T = \mathbb{N}$ or $T = \mathbb{Z}$. A stochastic process $X = (X_k)_{k\in T}$ has the Asymptotic Equipartition Property if $H(X)$ is finite and

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \text{ in probability as } n \to \infty.$$

Inpedendent and identically distributed processes with $H(X_1) < \infty$ have this property by Theorem 2.22 since $H(X) = H(X_1)$. But the Shannon-McMillan-Breiman theorem states that all stationary, ergodic processes with finite state space S have the AEP. We state the theorem now. Its rather long proof will be presented in the next section.

Theorem 2.24. *(The Shannon-McMillan-Breiman theorem) Let $X = (X_k)_{k\in\mathbb{Z}}$ be a stationary ergodic stochastic process taking values in a finite set S. If $H(X)$ is the finite entropy rate of the process, then*

$$\lim_{n\to\infty} -\frac{1}{n}\log p(X_0, X_1, \ldots, X_{n-1}) = H(X)$$

with probability 1 (and thus also in probability).

It is assumed that the process has parameter set $T = \mathbb{Z}$. But recall Theorem 0.22: any stationary stochastic process $X = (X_k)_{k\in\mathbb{N}}$ with parameter set $T = \mathbb{N}$ has an identically distributed counterpart process $X' = (X'_k)_{k\in\mathbb{Z}}$ with parameter set $T = \mathbb{Z}$.

Remark 2.25. The AEP actually holds for even wider class of processes. The state space may be countable, for instance. Moreover, if the X_k are continuous random variables and entropy is replaced with *differential entropy*, then the AEP again holds for stationary, ergodic processes [7]. But the proof of Shannon-McMillan-Breiman theorem in this case is considerably more difficult and way beyond the scope of this Master's Thesis.

The AEP is important because it enables us to divide the space of all sequences into typical and atypical sequences. This partitioning has important applications such as data compression, as we will soon see.

Definition 2.26. Let $T = \mathbb{N}$ or $T = \mathbb{Z}$. Suppose that $X = (X_k)_{k\in T}$ is a stochastic process with state space S, $H(X) < \infty$ and X has the AEP. Let $\epsilon > 0$. Then the *typical set* $A_\epsilon^{(n)}$ is the set consisting of sequences $(x_1, x_2, \ldots, x_n) \in S^n$ with the property

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}.$$

The following theorem shows that the probability of observing a sequence belonging to the typical set is close to 1, all elements of the typical set are approximately equiprobable, and the number of elements in the typical is close to $2^{nH(X)}$.

Theorem 2.27. *The set $A_\epsilon^{(n)}$ has the following properties:*

(1) *If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then $H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$,*

(2) *$P\left((X_1, X_2, \ldots, X_n) \in A_\epsilon^{(n)}\right) > 1 - \epsilon$ for large n,*

(3) *$|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$,*

(4) *$|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$.*

Proof. (1) This is immediate from the definition of $A_\epsilon^{(n)}$.

(2) Since X has the AEP, convergence in probability implies that for every $\delta > 0$ there exists $n \in \mathbb{N}$ such that

$$P\left(\left|-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) - H(X)\right| \leq \epsilon\right) > 1 - \delta. \tag{2.28}$$

If we choose $\delta = \epsilon$, then (2.28) says precisely that $P\left((X_1, X_2, \ldots, X_n) \in A_\epsilon^{(n)}\right) > 1-\epsilon$.

(3) Observe that

$$\begin{aligned} 1 &= \sum_{(x_1,x_2,\ldots,x_n)\in S^n} p(x_1, x_2, \ldots, x_n) \geq \sum_{(x_1,x_2,\ldots,x_n)\in A_\epsilon^{(n)}} p(x_1, x_2, \ldots, x_n) \\ &\geq \sum_{(x_1,x_2,\ldots,x_n)\in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} = |A_\epsilon^{(n)}|2^{-n(H(X)+\epsilon)}. \end{aligned}$$

The claim follows by dividing both sides by $2^{-n(H(X)+\epsilon)}$.

(4) For large n, we have $P\left((X_1, X_2, \ldots, X_n) \in A_\epsilon^{(n)}\right) > 1 - \epsilon$ by (2). Therefore,

$$\begin{aligned} 1 - \epsilon < P\left((X_1, X_2, \ldots, X_n) \in A_\epsilon^{(n)}\right) &= \sum_{(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}} p(x_1, x_2, \ldots, x_n) \\ &\leq \sum_{(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} = |A_\epsilon^{(n)}| 2^{-n(H(X)-\epsilon)}. \end{aligned}$$

The claim follows again by dividing both sides by $2^{-n(H(X)-\epsilon)}$.

□

Example 2.29. Suppose that $X_1, X_2, \ldots$ are independent Bernoulli(0.8)-distributed random variables as in the beginning of this section. If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) &\approx 2^{-nH(X)} = 2^{-nH(X_1)} = 2^{-nH(0.8)} = 2^{-n(-0.8 \log 0.8 - 0.2 \log 0.2)} \\ &= 0.8^{0.8n} 0.2^{0.2n}. \end{aligned}$$

Thus for typical sequences, around 80% of the X_k are ones. It is interesting that the most likely individual sequence, that is, the sequence in which every X_k is 1, does not belong to the typical set if ϵ is small enough. To see this, note that

$$-\frac{1}{n} \log p(1, 1, \ldots, 1) = -\frac{1}{n} \log 0.8^n = -\log 0.8 \approx 0.33 < 0.72 \approx H(0.8).$$

The following example illustrates why the AEP is useful.

Example 2.30. (Data Compression) Suppose that $X = (X_k)_{k \in \mathbb{N}}$ is a stochastic process with finite state space S, and suppose further that the AEP holds for X. Consider sequences $(x_1, x_2, \ldots, x_n) \in S^n$ drawn according to $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$. Since S^n has $|S|^n < \infty$ elements, these sequences can be represented with $\log |S|^n = n \log |S|$ bits (in practice, we of course need $\lceil n \log |S| \rceil$ bits since $n \log |S|$ may not be an integer). Let us call these bit representations *codewords*. By assigning shorter codewords to sequences that appear often and longer codewords to rare sequences, we can reduce the average codeword length. If $l(x_1, x_2, \ldots, x_n)$ is the length of the codeword associated with sequence $(x_1, x_2, \ldots, x_n)$, then the expected codeword length is

$$E[l(X_1, X_2, \ldots, X_n)] = \sum_{(x_1, x_2, \ldots, x_n) \in S^n} l(x_1, x_2, \ldots, x_n) p(x_1, x_2, \ldots, x_n).$$

Since there are at most $2^{n(H(X)+\epsilon)}$ sequences in $A_\epsilon^{(n)}$, we need no more than $n(H(X)+\epsilon)+1$ bits to represent each typical sequence (the one extra bit may be needed since $n(H(X)+\epsilon)$

may not be an integer). Let us then prefix these codewords with 0, so that no more than $n(H(X)+\epsilon)+2$ bits are needed to represent each sequence in $A_\epsilon^{(n)}$. Similarly, at most $n\log|S|+1$ bits are enough to represent all sequences not in $A_\epsilon^{(n)}$, and by prefixing these codewords with 1 we have maximum codeword length of $n\log|S|+2$ for sequences that belong to the complement of $A_\epsilon^{(n)}$.

Let n be so large that $P_{X_1,\ldots,X_n}\left(A_\epsilon^{(n)c}\right)$ is less than ϵ. Then

$$\begin{aligned}
E[l(X_1,\ldots,X_n)] &= \sum_{\mathbf{x}\in A_\epsilon^{(n)}} l(\mathbf{x})p(\mathbf{x}) + \sum_{\mathbf{x}\in A_\epsilon^{(n)}} l(\mathbf{x})p(\mathbf{x}) \\
&\leq \sum_{\mathbf{x}\in A_\epsilon^{(n)}} (n(H(X)+\epsilon)+2)p(\mathbf{x}) + \sum_{\mathbf{x}\in A_\epsilon^{(n)}} (n\log|S|+2)p(\mathbf{x}) \\
&= P_{X_1,\ldots,X_n}\left(A_\epsilon^{(n)}\right)(n(H(X)+\epsilon)+2) + P_{X_1,\ldots,X_n}\left(A_\epsilon^{(n)c}\right)(n\log|S|+2) \\
&\leq n(H(X)+\epsilon)+\epsilon n(\log|S|)+2 = n(H(X)+\epsilon'),
\end{aligned}$$

where $\epsilon' = \epsilon + \epsilon\log|S| + \frac{2}{n}$ can be made arbitrarily small. Therefore, on the average, sequences in S^n can be represented with $nH(X)$ bits. This is often considerably smaller than the $n\log|S|$ bits needed if codewords are assigned without taking advantage of the AEP. For example, in the coin tossing experiment we discussed in the beginning of this section, $H(X)$ is approximately 0.72, but $\log|S| = \log 2 = 1$.

Now it is time to finally prove the Shannon-McMillan-Breiman theorem.

2.4 Proof of the Shannon-McMillan-Breiman theorem

Our strategy is to show that with probability 1,

$$H(X) \leq \liminf_{n\to\infty} -\frac{1}{n}\log p(X_0,\ldots,X_{n-1}) \leq \limsup_{n\to\infty} -\frac{1}{n}\log p(X_0,\ldots,X_{n-1}) \leq H_k,$$

where H_k is a (nonrandom) sequence such that $H_k \to H(X)$ as $k\to\infty$. Of course this and the stationarity of X imply that

$$\lim_{n\to\infty} -\frac{1}{n}\log p(X_1,\ldots,X_n) = \lim_{n\to\infty} -\frac{1}{n}\log p(X_0,\ldots,X_{n-1}) = H(X),$$

and so the AEP holds for the process X. To achieve this goal, $-\frac{1}{n}\log p(X_0,\ldots,X_{n-1})$ will be "sandwiched" between two ergodic processes that converge to $H(X)$ and H_k, respectively.

The sequence H_k, called the kth-order entropy, is defined as

$$\begin{aligned} H_k &= E\left[-\log p(X_k|X_{k-1},\ldots,X_0)\right] = E\left[-\log p(X_0|X_{-1},\ldots,X_{-k})\right] \\ &= H(X_0 \mid X_{-1}, X_{-2}, \ldots, X_{-k}), \end{aligned}$$

where the second equation follows from stationarity. As in the proof of Theorem 2.20, stationarity and Theorem 2.17 imply that H_k is a decreasing sequence. We also have

$$\begin{aligned} \lim_{k\to\infty} H_k &= \lim_{k\to\infty} H_{k-1} = \lim_{k\to\infty} H(X_0 \mid X_{-1},\ldots,X_{-k+1}) = \lim_{k\to\infty} H(X_k \mid X_{k-1},\ldots,X_1) \\ &= H(X). \end{aligned}$$

To make the proof more comprehensible, it is divided into four steps.

(1) Define

$$H_\infty = E\left[-\log p(X_0 \mid X_{-1}, X_{-2}, \ldots)\right] = H(X_0 \mid X_{-1}, X_{-2}, \ldots).$$

In this step we apply martingale convergence theory to show that $\lim_{k\to\infty} H_k = H_\infty$. Since limits are unique, this further implies that $H(X) = H_\infty$.

First we prove that

$$H_k = E\left[-\sum_{x_0\in S} p\left(x_0 \mid X_{-1}, X_{-2}, \ldots, X_{-k}\right) \log p\left(x_0 \mid X_{-1}, X_{-2}, \ldots, X_{-k}\right)\right].$$

Put $g(x) = P(X_0 = x \mid X_{-1}, X_{-2}, \ldots, X_{-k}) = p(x \mid X_{-1}, X_{-2}, \ldots, X_{-k})$. Using the properties of conditional expectation, we obtain

$$\begin{aligned} H_k &= E\left[-\log p(X_0 \mid X_{-1},\ldots,X_{-k})\right] = E\left[-\sum_{x_0\in S} I_{\{X_0=x_0\}} \log g(x_0)\right] \\ &= -\sum_{x_0\in S} E\left[I_{\{X_0=x_0\}}\right] \log g(x_0) = -\sum_{x_0\in S} E\left[E[I_{\{X_0=x_0\}} \mid X_{-1},\ldots,X_{-k}]\right] \log g(x_0) \\ &= E\left[-\sum_{x_0\in S} E\left[I_{\{X_0=x_0\}} \mid X_{-1},\ldots,X_{-k}\right] \log g(x_0)\right] \\ &= E\left[-\sum_{x_0\in S} p\left(x_0 \mid X_{-1}, X_{-2}, \ldots, X_{-k}\right) \log p\left(x_0 \mid X_{-1}, X_{-2}, \ldots, X_{-k}\right)\right]. \end{aligned}$$

The same argument shows that

$$H_\infty = E\left[-\sum_{x_0\in S} p\left(x_0 \mid X_{-1}, X_{-2}, \ldots\right) \log p\left(x_0 \mid X_{-1}, X_{-2}, \ldots\right)\right].$$

Let $x_0 \in S$. Define $Y_k = p(x_0|X_{-1}, \ldots, X_{-k}), k \in \mathbb{N}$. Then by the definition of conditional probability,

$$Y_k = p(x_0|X_{-1}, \ldots, X_{-k}) = P(X_0 = x_0|X_{-1}, \ldots, X_{-k}) = E[I_{\{X_0=x_0\}}|X_{-1}, \ldots, X_{-k}].$$

As we saw in Example 0.42, the process $Y = (Y_k)_{k\in\mathbb{N}}$ is a martingale. We also observe that

$$\sigma\left(\bigcup_{n=1}^{\infty} \sigma(X_{-1}, \ldots, X_{-n})\right) = \sigma(X_{-1}, X_{-2}, \ldots).$$

Therefore, we may now apply Levy's martingale convergence theorem to obtain

$$\begin{aligned}\lim_{k\to\infty} Y_k &= \lim_{k\to\infty} E[I_{\{X_0=x_0\}}|X_{-1}, \ldots, X_{-k}] = E[I_{\{X_0=x_0\}}|X_{-1}, X_{-2}, \ldots] \\ &= P(X_0 = x_0|X_{-1}, X_{-2}, \ldots) = p(x_0|X_{-1}, X_{-2}, \ldots)\end{aligned}$$

with probability 1. Since $-x \log x \leq 1$ for $x \in [0, 1]$, we obtain

$$-\sum_{x_0\in S} p(x_0|X_{-1}, \ldots, X_{-k}) \log p(x_0|X_{-1}, \ldots, X_{-k}) \leq |S| < \infty$$

for all $k \in \mathbb{N}$. And since the function $-x \log x$ is also continuous on $[0, 1]$, we have $\lim_{k\to\infty} x_k \log x_k = (\lim_{k\to\infty} x_k) \log(\lim_{k\to\infty} x_k)$ for all convergent sequences $(x_k)_{k\in\mathbb{N}}$. Thus, the dominated convergence theorem yields

$$\begin{aligned}\lim_{k\to\infty} H_k &= \lim_{k\to\infty} E\left[-\sum_{x_0\in S} p(x_0|X_{-1}, \ldots, X_{-k}) \log p(x_0|X_{-1}, \ldots, X_{-k})\right] \\ &= E\left[-\sum_{x_0\in S} \lim_{k\to\infty} p(x_0|X_{-1}, \ldots, X_{-k}) \log p(x_0|X_{-1}, \ldots, X_{-k})\right] \\ &= E\left[-\sum_{x_0\in S} p(x_0|X_{-1}, X_{-2}, \ldots) \log p(x_0|X_{-1}, X_{-2}, \ldots)\right] = H_\infty.\end{aligned}$$

This completes the first part of the proof.

(2) The k-th order Markov approximation to the probability $p(X_0, X_1, \ldots, X_{n-1})$ is defined for $n \geq k$ as

$$p^k(X_0, X_1, \ldots, X_{n-1}) = p(X_0, X_1, \ldots, X_{k-1}) \prod_{i=k}^{n-1} p(X_i \mid X_{i-1}, X_{i-2}, \ldots, X_{i-k}).$$

In this step we use the ergodic theorem to prove that with probability 1,

$$\lim_{n\to\infty} -\frac{1}{n}\log p^k(X_0, X_1, \ldots, X_{n-1}) = H_k, \tag{2.31}$$

and

$$\lim_{n\to\infty} -\frac{1}{n}\log p(X_0, X_1, \ldots, X_{n-1} \mid X_{-1}, X_{-2}, \ldots) = H_\infty. \tag{2.32}$$

To prove (2.31), observe first that

$$-\frac{1}{n}\log p^k(X_0, \ldots, X_{n-1}) = -\frac{1}{n}\log p(X_0, \ldots, X_{k-1}) - \frac{1}{n}\sum_{i=k}^{n-1} \log p(X_i \mid X_{i-1}, \ldots, X_{i-k}).$$

The first term converges to zero as n grows, and the second term can be written as

$$-\frac{1}{n}\sum_{i=1}^{n-1} \log p(X_{i-1} \mid X_{i-2}, \ldots, X_{i-k-1}) + \frac{1}{n}\sum_{i=1}^{k-1} \log p(X_i \mid X_{i-1}, \ldots, X_{i-k}),$$

where the second term again converges to zero as n grows. The ergodic theorem can be applied to the first term, since

$$\log p(X_{i-1} \mid X_{i-2}, \ldots, X_{i-k-1}) = f\left[T^{i-1}(\ldots, X_{-1}, X_0, X_1, \ldots)\right],$$

where $f(\ldots, x_{-1}, x_0, x_1, \ldots) = \log p(x_0 \mid x_{-1} \ldots x_{i-k})$, and T is the shift operator on $S^{\mathbb{Z}}$. The function f is measurable by the definition of conditional probability, and since

$$-E[f] = E\left[-\log p(X_0 \mid X_{-1}, \ldots, X_{-k})\right] = H_k < \infty,$$

it is integrable. Therefore, by (1.28), we have

$$\begin{aligned}\lim_{n\to\infty} -\frac{1}{n}\log p^k(X_{n-1}, \ldots, X_0) &= \lim_{n\to\infty} -\frac{n-1}{n(n-1)}\sum_{i=1}^{n-1} \log p(X_{i-1} \mid X_{i-2}, \ldots, X_{i-k-1}) \\ &= -E\left[f(\ldots, X_{-1}, X_0, X_1, \ldots)\right] \\ &= -E\left[\log p(X_0 \mid X_{-1}, \ldots, X_{i-k})\right] = H_k.\end{aligned}$$

To prove (2.32), recall that conditional probability satisfies the equation

$$P(X = x, Y = y \mid Z = z) = P(Y = y \mid X = x, Z = z)P(X = x \mid Z = z)$$

even if $P(Z = z) = 0$ (see Example 0.37). Therefore,

$$\begin{aligned} p(X_0, \ldots, X_{n-1} \mid X_0, X_1, \ldots) &= p(X_1, \ldots, X_{n-1} \mid X_0, X_{-1}, \ldots) p(X_0 \mid X_{-1}, X_{-2}, \ldots) \\ &= \cdots = \prod_{i=0}^{n-1} p(X_i \mid X_{i-1}, X_{i-2}, \ldots) \end{aligned}$$

with probability 1. This implies that

$$\begin{aligned} -\frac{1}{n} \log p(X_0, \ldots, X_{n-1} \mid X_{-1}, X_{-2}, \ldots) &= -\frac{1}{n} \sum_{i=0}^{n-1} \log p(X_i \mid X_{i-1}, X_{i-2}, \ldots) \\ &= -\frac{1}{n} \sum_{i=1}^{n} g\left[T^{i-1}(\ldots, X_{-1}, X_0, X_1, \ldots)\right], \end{aligned}$$

where $g(\ldots, x_{-1}, x_0, x_1, \ldots) = \log p(x_0 \mid x_{-1}, x_{-2}, \ldots)$, and T is the shift operator on $S^{\mathbb{Z}}$. The function g is again measurable and integrable, and thus the ergodic theorem implies that

$$\lim_{n\to\infty} -\frac{1}{n} \log p(X_0, \ldots, X_{n-1} \mid X_{-1}, X_{i-2}, \ldots) = -E\left[g(\ldots, X_{-1}, X_0, X_1, \ldots)\right] = H_\infty.$$

This completes the second part of the proof.

(3) In this part we prove two limit inequalities, namely

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{p^k(X_0, X_1, \ldots, X_{n-1})}{p(X_0, X_1, \ldots, X_{n-1})} \leq 0 \tag{2.33}$$

almost surely, and

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{p(X_0, X_1, \ldots, X_{n-1})}{p(X_0, X_1, \ldots, X_{n-1} \mid X_{-1}, X_{-2}, \ldots)} \leq 0 \tag{2.34}$$

almost surely. Observe that since $(X_0, X_1, \ldots, X_{n-1})$ is a discrete random variable, division by zero is not an issue here.

Let A be the support set of $P_{X_0, X_1, \ldots, X_{n-1}}$, that is,

$$A = \{(x_0, x_1, \ldots, x_{n-1}) \in S^n : p(x_0, x_1, \ldots, x_{n-1}) > 0\}.$$

Then

$$
\begin{aligned}
E\left[\frac{p^k(X_0,X_1,\ldots,X_{n-1})}{p(X_0,X_1,\ldots,X_{n-1})}\right] &= \sum_{(x_0,\ldots,x_{n-1})\in A} p(x_0,\ldots,x_{n-1})\frac{p^k(x_0,\ldots,x_{n-1})}{p(x_0,\ldots,x_{n-1})} \\
&= \sum_{(x_0,\ldots,x_{n-1})\in A} p^k(x_0,\ldots,x_{n-1}) \le \sum_{(x_0,\ldots,x_{n-1})\in S^n} p^k(x_0,\ldots,x_{n-1}) \\
&= \sum_{(x_0,\ldots,x_{n-1})\in S^n} p(x_0,\ldots,x_{k-1})\prod_{i=k}^{n-1} p(x_i \mid x_{i-1},\ldots,x_{i-k}) \\
&= \sum_{(x_0,\ldots,x_{n-2})\in S^{n-1}} p(x_0,\ldots,x_{k-1})\prod_{i=k}^{n-2} p(x_i \mid x_{i-1},\ldots,x_{i-k}) \underbrace{\sum_{x_{n-1}\in S} p(x_{n-1} \mid x_{n-1-1},\ldots,x_{n-1-k})}_{=1} \\
&= \sum_{(x_0,\ldots,x_{n-2})\in S^{n-1}} p(x_0,\ldots,x_{k-1})\prod_{i=k}^{n-2} p(x_i \mid x_{i-1},\ldots,x_{i-k}) = \cdots = \sum_{(x_0,\ldots,x_{k-1})\in S^k} p(x_0,\ldots,x_{k-1}) \\
&= 1.
\end{aligned}
$$

Thus, by Chebyshev's inequality, we have

$$
P\left(\frac{p^k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1}\mid X_{-1},X_{-2},\ldots)} \ge n^2\right) \le \frac{E\left[\frac{p^k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1}|X_{-1},X_{-2},\ldots)}\right]}{n^2} \le \frac{1}{n^2}
$$

or equivalently,

$$
P\left(\frac{1}{n}\log\frac{p^k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1}\mid X_{-1},X_{-2},\ldots)} \ge \frac{1}{n}\log\frac{1}{n^2}\right) \le \frac{1}{n^2}.
$$

Since the series $\sum_{n=1}^{\infty}\frac{1}{n^2}$ converges to a finite number, the Borel-Cantelli lemma implies that with probability 1 the event

$$
\left\{\frac{1}{n}\log\frac{p^k(X_0,\ldots,X_{n-1})}{p(X_0,\ldots,X_{n-1}\mid X_{-1},X_{-2},\ldots)} \ge \frac{1}{n}\log\frac{1}{n^2}\right\}
$$

occurs only finitely many times. But since

$$
\lim_{n\to\infty}\frac{1}{n}\log\frac{1}{n^2} = \lim_{n\to\infty} 2\left(\frac{1}{n}\log\frac{1}{n}\right) = 0,
$$

this clearly implies that (2.33) holds with probability 1.

Let $(x_{-1}, x_{-2}, \ldots) \in S^\infty$, and put

$$\begin{aligned} g(x_{-1}, x_{-2}, \ldots) &= E\left[\frac{p(X_0, \ldots, X_{n-1})}{p(X_0, \ldots, X_{n-1} \mid X_{-1}, X_{-2}, \ldots)}\middle| X_{-1} = x_{-1}, X_{-2} = x_{-2}, \ldots\right] \\ &= E\left[\frac{p(X_0, \ldots, X_{n-1})}{p(X_0, \ldots, X_{n-1} \mid x_{-1}, x_{-2}, \ldots)}\middle| X_{-1} = x_{-1}, X_{-2} = x_{-2}, \ldots\right]. \end{aligned}$$

Also, let $B(x_{-1}, x_{-2}, \ldots) \subset S^n$ be the support set of $p(x_0, \ldots, x_{n-1} \mid x_{-1}, x_{-2}, \ldots)$, that is,

$$B(x_{-1}, x_{-2}, \ldots) = \{(x_0, \ldots, x_{n-1}) \in S^n : p(x_0, \ldots, x_{n-1} \mid x_{-1}, x_{-2}, \ldots) > 0\}.$$

Then

$$\begin{aligned} g(x_{-1}, x_{-2}, \ldots) &= \sum_{(x_0, \ldots, x_{n-1}) \in B(x_{-1}, \ldots)} p(x_0, \ldots, x_{n-1} \mid x_{-1}, \ldots) \frac{p(x_0, \ldots, x_{n-1})}{p(x_0, \ldots, x_{n-1} \mid x_{-1}, x_{-2}, \ldots)} \\ &= \sum_{(x_0, \ldots, x_{n-1}) \in B(x_{-1}, \ldots)} p(x_0, \ldots, x_{n-1}) \leq 1. \end{aligned}$$

Therefore, using the law of iterated expection, we obtain

$$\begin{aligned} E\left[\frac{p(X_0, \ldots, X_{n-1})}{p(X_0, \ldots, X_{n-1} \mid X_{-1}, X_{-2}, \ldots)}\right] &= E\left[E\left[\frac{p(X_0, \ldots, X_{n-1})}{p(X_0, \ldots, X_{n-1} \mid X_{-1}, X_{-2}, \ldots)}\middle| X_{-1}, \ldots\right]\right] \\ &= E\left[g(X_{-1}, X_{-2}, \ldots)\right] \leq 1. \end{aligned}$$

From here, (2.34) follows again by applying the Chebyshev inequality and the Borel-Cantelli lemma. This completes the third part of the proof.

(4) This is the final part of the proof. First, the inequality (2.33) of part 3 and equation (2.31) of part 2 imply that

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{p(X_0, X_1, \ldots, X_{n-1})} \leq \limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{p^k(X_0, X_1, \ldots, X_{n-1})} = H_k$$

with probability 1. Similarly, (2.34) and (2.32) imply that

$$\liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{p(X_0, \ldots, X_{n-1})} \geq \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{p(X_0, \ldots, X_{n-1} \mid X_{-1}, \ldots)} = H_\infty$$

with probability 1. Putting these inequalities together, we obtain

$$H_\infty \leq \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{p(X_0, \ldots, X_{n-1})} \leq \limsup_{n\to\infty} \frac{1}{n} \log \frac{1}{p(X_0, X_1, \ldots, X_{n-1})} \leq H_k$$

almost surely. We proved in part 1 that $\lim_{k\to\infty} H_k = H_\infty$. Therefore

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{p(X_0, \ldots, X_{n-1})} = H_\infty = H(X) \text{ (a.s.)}$$

which completes the proof.

□

Bibliography

[1] Ash Robert B., Doleans-Dade Catherine: Probability & Measure Theory, Second Edition, Academic Press, 2000.

[2] Ash Robert B.: Information Theory, Dover, 1990.

[3] Billingsley Patrick: Probability and Measure, Third Edition, Wiley Interscience, 1995.

[4] Cover Thomas M., Thomas Joy A.: Elements of Information Theory, Second Edition, Wiley Interscience, 2006.

[5] Ihara Shunsuke: Information Theory, World Scientific Publishing Company, 1992.

[6] Rosenthal Jeffrey S. : A First Look at Rigorous Probability Theory, Second Edition, World Scientific Publishing Company, 2006.

[7] Algoet Paul, Cover Thomas M., A sandwich proof of the Shannon-McMillan-Breiman theorem, Ann. Prob., 16(2): 899-909, 1988.

[8] Yeh J., Real Analysis: Theory of Measure and Integration, Second Edition, World Scientific Publishing Company, 2006.